BU-48-M RHD Stat - Sept 1953

Not for Distribution

PLANT BREEDING 210

CHAPTER I

INTRODUCTION

Summary. In this chapter, attention is called to the prevalence of statistics in the everyday activities of individuals and to statistics as a part of scientific method. In addition some history is presented. It is hoped that this material will orient the reader and provide motivation other than that of necessity.

Everyday statistics. Statistical literacy is already a necessity for effective citizenship. It is an unusual person who does not encounter and use statistics daily. Consider the daily paper. All but the youngest members of the family read it, In the appropriate seasons it contains information on batting averages, passes completed or intercepted in football, weights of hockey team members, heights of basketball players and a host of other statistical information on sports and players. Statistical misinformation is also available and presented with apparent authority. Often the purpose of such information is extraction of dollars from the unwary but the method may vary from crude to subtle. As holiday seasons approach, editorials present accident statistics in an attempt to prevent further accidents. The financial section may be devoted entirely to statistics and their interpretation. Other common statistics are the average weights for specified heights available from penny scales.

The individual goes through a random sampling procedure in the tossing of a coin. Upon the result, a decision between two courses of action of apparently equal merit is made. Our opinion of a new product may be requested in order that a manufacturer may obtain information which will enable him to decide whether or not to place it on the market. Thus, the individual supplies an observation for a sample; if you like, he is sampled.

In addition to being confronted daily with statistics, to sampling and being sampled, the individual makes probability statements of the sort that statisticians make when they wish to be precise yet, at the same time, leave themselves a loophole. Thus, if an individual states that "It will probably rain before noon", he means that it is more likely to rain before noon than not in his estimation. If the individual who made the above statement is prepared to lay a bet on the outcome of the morning's weather, such that he is prepared to win or lose in any individual instance but expects neither to

lose nor win in the long run, then he is also a statistician. If he will put up money only equal to that of his companion, then he is guessing but if he is prepared to wager more than his companion, presumably he is using knowledge, evidence, and reason to make a statistical inference. It must, of course, be kept in mind that "fools for arguments use wagers". Whenever words such as probably and likely are used, a statement is being made that is approximately statistical.

Statistics. What do statisticians do? Statisticians, and we shall not attempt to define the word statistician, deal with the collection of data, the summarization of data to obtain and present certain facts that characterize them, and the making of inferences more widely applicable than to the collected data themselves. Some statisticians deal solely with the mathematical aspects of the problems of statistics while others are solely concerned with data and the application of statistical techniques to their collection, characterization and interpretation. These individuals represent the extremes. This book is for the beginner in statistics who plans to be a user of statistical methods. Since the mathematical computations will require but little in the way of intelligence, if the reader has the desire to perform them it is hoped that he will use his time to gain an insight and understanding of the subject rather than simply to acquaint himself with the recipes contained on the pages.

The statistician collects data. In this task, there is a wide range of activity. The census taker is a collector of data. A census is the enumeration of all the data of a certain type as, for example, a census of farm properties for purposes of taxation. On the other hand, a pollster may ask relatively few people a question designed to elicit information relative to their voting preferences. Very different uses of these data may be planned. The collection and classification of facts is often simply a record of history and at other times the basis of inference and future action. The emphasis has shifted in the quarter century since the publication of R. A. Fisher's Statistical Methods for Research Workers from the past to the future. It is the forward looking process with which the authors of this text wish to deal more particularly.

The statistician summarizes data. Information contained in masses of data is generally difficult to comprehend. Consequently, it is desirable to present certain representative facts contained in the data in abbreviated form. A birthrate has more meaning for most individuals than a statement of the number of childless couples, number of couples with one child, two children, etc.,

although the latter statements contain much more information. The latter statements, however, are also summaries of the more complete information which states which individuals have a specified number of children. Rates, ratios, means, medians, variances, standard deviations, pictograms and graphs are among the quantities and devices which the statistician uses to present data in compact, easily intelligible form. These and others together with the circumstances under which each is appropriate will be considered in this text.

The statistician makes inferences. Making inferences, certain or uncertain, is a practice of most people. With statisticians, it's business. Statements such as "It will probably rain by evening", "The Toronto Maple Leafs will probably win the Stanley Cup this winter", and "I'm not likely to win the office football pool" are examples of uncertain inference. The statistician's contribution to uncertain inference is the assignment of a measure of the uncertainty. Thus, for example, a statistician might consider the available data on time of latest spring frost for a locality and make a prediction or uncertain inference of the following sort. In the coming spring, the odds are even that the latest spring frost will not occur after May 16. Since these odds may not be very helpful to a farmer, the statistician may wish to go on and say that unless an unusual year is in store, say one likely to occur only once every 20 years, then the time of latest spring frost should be prior to June 3 or whatever date his calculations yield. The statistician will always state a measure of the uncertainty of his inference in a statement such as "unless an event such as is likely to occur only once in (e.g., twenty) times has occurred". A statistical proof is always <u>proof beyond a reasonable doubt</u> rather than an incontrovertible proof.

This book is intended to give the student the necessary background to make his own inferences and view other persons' inferences critically. Too often, figures are presented in the hope that no one will have the temerity to dispute them. Unfortunately, such presentation is, at times, with enthusiasm and conviction where hesitation and caution are more to be desired. The uninitiated develop a feeling of inferiority and resentment. It is hoped that this text can dispel these feelings from the minds of its readers.

<u>Biometry</u>. This text is aimed, in particular, at those in the biological sciences who wish to apply statistics. Biometry is statistics applied to the biological sciences. It is more nearly quantitative biology than mathematical biology and requires cooperation among biologists, mathematicians and statisticians for its development. Biometricians tend to think of themselves in rela-

tion to the common interests of the three groups rather than in terms of a special discipline although most have had their training largely in one discipline. Clearly, all statisticians are not biometricians.

Most workers in biological fields sooner or later possess data which require statistical treatment. Few such workers have sufficient time to become competent statisticians. A purpose of this text is to enable such workers to recognize the need, when it is present, for the cooperation of a biometrician or statistician in the collection, summarization and interpretation of their data. Such cooperation should always be sought prior to the collection of data. If this is done, much disappointment can often be avoided.

History. It would seem pertinent in our introductory chapter to touch briefly on the historical development of statistics and biometry. Initially, statistics would appear to be the arithmetic of the state. Rulers required information about their subjects in order that they could levy taxes that would have a reasonable hope of being collected or in order that they would have sufficient reserves of men and material for waging war.

Probably all cultures that intentionally recorded their own history, recorded certain statistics concerning themselves. In some cases the statistics are available, while in others the recording seems to have had more effect upon history than the record. For example, it is recorded that Caesar Augustus sent out a decree that all the world should be taxed. Statisticians were available in Bethlehem rather than Nazareth and, as a result, Jesus was born at the former place though his parents must have claimed Nazareth as their home. The statistics obtained are not available as far as the authors know but the name of Bethlehem has become certainly as commonplace as that of Nazareth. The statistician seems to have affected history in the collection of his data rather than as a result of the collection.

Ship insurance seems to have been available in the fourteenth century to Flemish shipping, at least. This sort of financial speculation must have been sheer gambling in its initial stages. However, insurance in general is now so reputable that not to carry a variety of policies is considered a mark of imprudence. Life insurance is based on mortality tables, tables prepared from experience and giving information on numbers living and dying in the different age groups. For example, a common table takes 100,000 people at age 10, gives the number dying in each year, and computes from this information the number living at each age, and the yearly probabilities of living and dying. Similar tables are required for other types of insurance. Clearly, there's work here

for the statistician.

It was the more formal games of chance, such as those involving cards and dice, that appear to have given rise to the modern theory of probability. Apparently the astrologer was insufficient to the task though he is still able to make a living. About the middle of the seventeenth century, Pascal was presented with certain questions arising out of the gambling experiences of the Chevalier de Méré. Pascal and Fermat had some correspondence on the subject and started the theory of probability. Toward the end of the same century, Halley published a Life Table applicable to the city of Breslau in the Philosophical Transactions of the Royal Society of London. Probability and insurance have continued the close relationship.

De Moivre in 1733 published the equation of the normal curve. Laplace and Gauss, neither of whom had been born at the time, derived the same curve of error independently of each other and of de Moivre. This curve and others derived from it are basic in modern statistics. De Moivre had no notion of applying his law to statistical data and the other two used their findings in the field of mathematical astronomy. Quetelet, a Belgian astronomer, believed that application was valid to data from biological, social and physical sciences. All this was prior to the 20th century.

During this early period many other mathematicians, both well-known and less well-known, contributed to the theory of probability. Much of their work dealt with games of chance but some were interested in determining population growth, annuities, sex ratios and probabilities of human life.

The need of the application of mathematics to the biological sciences was probably spurred by the work of the biologist, Darwin, 1809-1882, who infused biology with a new enthusiasm. Among those first to apply mathematics were Sir Frances Galton and Karl Pearson. Much of their work falls in the category of large sample theory.

One of Pearson's students was William Sealy Gosset (1876-1937), commonly known as "Student". In 1908, he published his t-test which was particularly applicable to hiw work involving small samples. Pearson dealt with samples so large that the refinement available in the t-test was of little practical importance. This test was the first contribution of statistics to be uniformly useful to the biologist. Briefly, the test permits us to assign probabilities to the deviations of the means of random samples from populations having having a known or hypothesized mean. For example, suppose a certain feeding program for fowl results in a bird of x pounds, on the average, at market time.

Company A declares that a feeding program using their feed will result in a bird of x+y pounds, on the average, at market time. Since y pounds increase sounds as though it were economically advantageous, chicken farmer F decides to put 20 birds on the program and give it a tryout. Obviously, farmer F won't get exactly x+y pounds as the average of his twenty birds. The question is how much leeway can be tolerated before the farmer must declare company A has been overenthusiastic or underenthusiastic about its product. In a later chapter, we shall see how the t-test answers this question.

Sir Ronald A. Fisher (1890 -) developed statistics further and to such an extent that it is quite fair to say that these developments constitute a revolution in methods of experimentation. Some of his ideas were slow to be adopted, even in his own country, but they have been adopted within his lifetime and the emphasis on the past, prevalent throughout statistics up to the twentieth century, has shifted to the future, since publication of his Statistical Methods for Research Workers.

Currently, many people are contributing to the field of statistics. It is a new field and one in which the concepts are still fluid. The basic problems have to do with making decisions and with the consequences of these decisions. The statistician relates these to probabilities. It is clear, then, that the basic problems are not peculiar to the subject matter fields. However, we shall draw the large part of our illustrative material from biology and related fields.

The scientific method. In the previous section we stated that R. A. Fisher revolutionized the methods of experimentation. An experiment is a trial or the making of special observations for the purpose of confirming or disproving something about which the experimenter is doubtful. For the scientific experimenter, this "something" will usually be a suggested truth or hypothesis. His problem is to check on its validity in a scientific manner.

The scientific method of experimentation involves three steps taken in the following order: i) the formulation of an hypothesis, ii) the conduct of an experiment, and iii) the drawing of a conclusion. We shall see that statistics is an integral part of the scientific method and not simply an adjunct to it.

An hypothesis is generally formulated on the basis of the observations of the experimenter or others. There is doubt as to the validity of the hypothesis since the observations were, in general, not obtained for the purpose of testing such an hypothesis. Consequently there are alternatives to the

hypothesis and these may range in precision of statement from a single specific alternative to one so vague that it is little more than a statement that the proposed hypothesis is not necessarily correct.

A typical experiment is performed with a limited quantity of experimental material which is considered to be representative of the main body of material. Clearly, this experimental material will not give exactly the results to be expected from the main body itself. One must then make an inference, an uncertain inference, about the main body from the facts obtained in the experiment.

In making this uncertain inference, the experimenter is taking cognizance of the fact that he is dealing with a material and not an ideal world. He appreciates that he may be in error but does his best to make the possibility of an error small. Now the function of statistics is to evaluate numerically the uncertainty of the inference. Having a number that measures the uncertainty of his inference, the experimenter has a measure of the risk involved in the decision he makes. This decision may affect his professional reputation, his employer's profit or loss, or the lives of the sick or injured. The choice of a standard of such risk in making decisions is that of the man making the decision, not that of the statistician.

The process that leads to an uncertain inference is one of induction. Prior to the twentieth century, this principle was not well developed and for this reason, statistics did little more than record history. Induction is the process of reasoning from particulars to generals, from the individual to the universal, or as the statistician says, from the sample to the universe or population. Statisticians are, necessarily, inductive thinkers. On the other hand, the mathematician uses induction comparatively rarely and when he does he begins with a demonstration that a certain law holds in a specific case, proves that if it holds in a certain case it must hold in the next, and therefore in the next, and so ad infinitum.

The mathematician is a deductive thinker. He begins with a set of elements, for example, the integers, defines for himself certain operations, for example, addition and multiplication, and goes on to draw inferences that are indisputable conclusions and not subject to uncertainty. The principle of deduction has been well developed for centuries and, consequently, mathematics is a much older field than statistics. Statistics uses much of mathematics but the modes of thought are different: mathematics deals with an ideal world while statistics deals with a material world.

Variation. Since the statistician deals with a limited quantity of exper-

imental material and from this must make an inference about the main body, he must carefully consider why the experimental material does not exactly depict the response of the whole. This characteristic lack of uniformity of outcome is generally referred to as variation. Apart from certain major sources of variation, there are two which give the experimenter most concern. These are inaccuracy and error. We associate inaccuracy with the individual or the measuring device and the experimenter can often do something about this. In any case, he tries to control it. The term error is introduced at this stage because it is used throughout statistics, usually qualified by the word experimental. The term has no adverse connotations. Variation is inherent in the experimental material. The statistician has converted this apparently troublesome characteristic into a yardstick to be used in measuring departures from expected or hypothesized outcomes for the purpose of making a decision about the validity of the hypothesis. For this purpose, variation is first converted to a number which is called experimental error. Inaccuracy and error, not distinguishable in single observations, vary in magnitude from experiment to experiment within a field and from field to field. In biological and agricultural research, experimental error tends to be much larger while in chemistry and physics, error of measurement is often the difficult error to control. In any case, variation is universal. In a later chapter, it will be shown how some inaccuratenesses can be relegated to error by a process called randomization.

Designing experiments. Experiments are performed to obtain information from which inferences may be made. In designing an experiment, the first thing to be considered is the clear definition of the objectives. An ill-designed experiment may be found to throw no light upon the objectives the experimenter had in mind initially. A clear simple statement of the hypotheses to be tested and the alternatives to be accepted if the data do not support them is the usual form of such definition. Next the available experimental material, labor, money, etc., must be taken into consideration since if the experiment cannot be carried out adequately and competently, the experimenter's energies are better directed elsewhere. In some cases, the result may be a smaller experiment with less widely applicable inferences.

Many possible designs are available for experimentation. The names randomized blocks and Latin squares must be familiar to many with no training in statistics. The choice of a design is made on the basis of the criterion that the design be simple analytically yet control the major sources of

experimental variation such as type of soil, breed of cattle, day of the week or whatever is applicable. The size of the experiment is a part of its design and the well-designed experiment uses a minimum of experimental material while yielding conclusive results. The determination of the size of an experiment is based on the amount of variation to be expected in the experimental material and upon the risks involved in making the decisions arising from acceptance or rejection of the various hypotheses.

Warning. By our enthusiasm about statistics, we hope that we have not made the subject appear either excessively difficult or mysterious. It is unfortunate that some users of statistics have made it appear capable of obtaining information from the most inadequate data. A moment's thought will convince the reader that only that information which the experiment was designed to obtain, can be obtained from the resulting data; and then only if the experiment has been capably conducted throughout.

Scope of the text. This is an elementary non-mathematical text. It is the intention of the authors to give the reader an understanding of those modern statistical concepts that the experimenter most often uses in making decisions affecting his future actions. Principles rather than techniques are presented, although the techniques are there to exemplify the principles. It is hoped that the examples are simple enough that the student will not spend an undue amount of time on them but that they will aid him in grasping the principles securely. It is hoped that after study of this text, the student will be able to appreciate those papers with statistical content that become a part of his reading and that he will feel capable of using those valuable texts that contain so many of the experimental designs that are used in the conduct of research. Finally, the authors wish to leave with the reader enough knowledge of statistics that he will appreciate that there are times when he should consult a statistician and that he will be able to converse intelligently with the statistician in complete cooperation.

CHAPTER II

OBSERVATIONS

2.1 Summary. This chapter contains a discussion of variables, distributions, populations, parameters, samples and statistics. The relation of these terms to everyday realities and less technical usage is discussed. The collection and summarization of data are discussed and the making of inferences, in the form of confidence interval statements, is introduced.

2.2 Variable and Variate. We commonly make statements to distinguish between objects. For example, we say that Helen is blonde, that the horse we admire is especially large, that we prefer skim milk, and so on. These statements do make a distinction because the characteristics, color, size, and creaminess, are variable ones. If they were not, no distinction would be possible. Thus, we define a random variable as any characteristic of an individual which shows variation from individual to individual within the group we have in mind.

We also define a variate as the particular value taken by a variable, i.e. a variate is the individual observation. The variates above were blonde. large, and skim.

Instead of writing the words variable and variate each time, we shall use the following shorthand notation: X to denote the variable under study and X_i to denote the i-th observation. In flipping a coin ten times, we may note the number of heads. The variable X is the number of heads occurring in ten tosses. If we perform the operation 3 times and obtain 6, 3. and 8 heads, then the variates are $X_1 = 6$, $X_2 = 3$, and $X_3 = 8$. If we wish to speak very generally, we simply use $X_1, X_2, \ldots, X_n$. In the coin-tossing example, n equalled 3.

The usual experimental material that is dealt with has many variables. In our early chapters, we will be concerned with the study of one variable at a time.

Variables may be classified in several ways. A variable may be either qualitative or quantitative. Flower color is an example of a qualitative variable whereas weight in pounds is a quantitative variable. Variables are also classified as being either continuous or discontinuous (discrete). A continuous variable is one in which the variates can assume any value within the total range of the population but, of course, the number of possible values that can be recorded is limited by the refinement of measurement used. Height is a continuous variable, whereas the number of germinating seeds in a sample of 100 is a discontinuous or discrete variable. For the present we will be concerned primarily with data

which are continuous. Such data are frequently referred to as measurement data.

2.3. The distribution of a variable. It has been said that variables are used to describe and distinguish among the individuals in some group. Actually, most of us manage to do a bit better than simply use a variable because we have ideas about the values that variables assume. In Minnesota, blondness is not a rare characteristic; most people would not put much credence in stories about a 65 lb. cat; but the statement that Henry weighs 165 lbs. would raise no eyebrows. Something is being implied, at least, about possible values and their frequencies in some group to which the individuals belong. We shall be a bit more formal and say that a certain variable is distributed in a specific fashion. The variable, then, has a distribution. For example, if a well-balanced arrow is spun on a dial having 10 equal divisions, there is no reason to expect that the arrow will prefer to stop on one or more numbers more frequently than on the others. The variable is the number appearing at the pointer's end, and it is said to be uniformly distributed or to have a uniform distribution.

A simple example of a distribution has just been given. What can be done about more complicated ones? When the experimenter has some data and experience that tell him what sort of a distribution he is dealing with, he can take his problem to a mathematician or statistician who may be able to suggest some mathematical expression to describe the manner in which the data are distributed. Together, they can usually reach a decision. The mathematics and statistics can then be developed. It will not be unusual if the mathematical expression makes provision for some events which the experimenter knows to be impossible. However, if the mathematical expression indicates that these events are likely to occur only once in an extremely large number of occurrences, say once in 19 billion times, the experimenter should be very grateful for the help received. Further experience will throw light on how good an approximation to reality the mathematical expression is.

The experimenter is often prepared to ignore the mathematical expression for the distribution of his variable provided he is given some characteristic quantity or quantities that give him sufficient information. Such quantities are called parameters. These are the constants of the distribution. They are generally unknown, and one of our problems will be to estimate them and make inferences about them. Particular examples will be discuased in the next chapter.

2.4 Samples and statistics. The statistician is prepared to use the data obtained by random sampling procedures to make inferences. When data are obtained by non-random sampling procedures, one does not feel that the inferences are necessarily valid and it is not possible to assign any number as a measure of the uncertainty of the inference. Randomness must be built into an experiment by a mechanical operation and is not to be confused with haphazardness. It is a means of assuring that the natural variation of a characteristic supplies us with a unit that can inform us, with validity, how large a departure from our hypothesis can be expected to be reasonable.

In general, all the information in a sample is difficult to grasp. If a coin is flipped 10 times, it is customary to accept the summarized information available in the number of heads obtained rather than to insist upon knowing which of the ten trials resulted in heads. It is customary, then, to summarize the collected data. The numerical bits of information that summarize the data are referred to as statistics; means and variances are two common statistics which will be discussed continuously throughout the text. The statistics of a sample have their counterparts in the parameters of the population. Statistics are usually estimates of parameters. It is to be noted that statistics are variables and not constants. Thus if the mean of a sample is calculated, a variate, i.e., a number, is obtained but the sample is only one of many possible samples, in general.

Statistics, then, are variables and it is possible to determine their distribution when the distribution of the individuals is known. A knowledge of the distribution of a statistic is required before one can make an inference. Parameters are constants and, as such, do not have distributions.

A sampling exercise. To show how the distribution of a statistic can be used to make an inference, consider Table 2.1, a table of random numbers. This table consists of the numbers 0, 1, 2, ..., 9 arranged in a 100 by 100 table, i.e., there are 10,000 of these numbers. These numbers were turned out by a machine set up to do the job in a random fashion. There was no reason to expect it to turn out any one number more often than another nor any sequence of numbers more often than others, except by chance. You will find that there are 0's, 1's, ..., 9's, and that are odd. Draw samples of size 10 from this table, ignoring the individual numbers and simply recording the number of odds present, till a large number of samples has been obtained.

You may start anywhere in the table, but a very satisfactory way is to stab at one of the pages, read the four numbers most nearly opposite the top of your finger, and use these numbers to locate a sampling point. Thus, if

the four numbers turn out to be 3384, begin your first sample at the intersection of row 33 and column 84. Proceed from this point in any direction till you have your first sample of 10 digits. If you proceed down from this point, your sample is 48 75995 648. You can obtain a lot of such samples in a very short time. A large class obtained the following results:

Table 2.2

With such a large number of samples, we feel fairly confident that Table 2.2 is representative of the results we would obtain if we took all possible samples. Actually, that statement need not call upon our intuition as the mathematicians have a theorem which states, in more precise terms, the same result.

The table indicates that samples with 0, 1, 9, and 10 odd numbers are unusual while samples with 4, 5, and 6 odd numbers are quite common. The other numbers of odds lie in between in frequency. Notice that approximately 95% of all samples have from 2 to 8 odd numbers, while approximately 99% have 1 to 9 odd numbers.

Now suppose that a sample has been obtained from a population in which the percentage of odd numbers is unknown but is hypothesized as being 50%. Instead of odd numbers, the problem can be considered as one of flipping a coin, calling a head odd and a tail even; or it can be the asking of a question with possible answers "yes" and "no" expected to occur with equal frequency. How far from 50% can we reasonably expect our sample to be? To answer this, decide how often you can afford to be wrong when 50% is the right answer. Naturally you want to be right all of the time, but you are

sampling and a sample is only a part of the whole. You must make an uncertain inference. Can you afford to say 50% is the wrong population percentage when it is really the correct one, one time in 20 such trials? If so, you can accept anywhere from 2 to 8, inclusive, odd numbers in your sample.

Other tables such as 2.2 can be built up from the same table of random numbers. If a 2:1 ratio is hypothesized, divide the random numbers obtained by 3 and observe the remainder. The resulting numbers, 0, 1, and 2, do not occur with equal frequency unless you discard one, say 0 or 9, yielding a zero remainder. Now you have used all but 10% of the numbers to give the numbers 0, 1, and 2 and can say that 0 and 1 represent a certain characteristic and 2 represents its absence. Draw your samples of the desired size and build up a table similar to 2.2 for the ratio 2:1 in this empirical manner. This table can be used in a manner similar to that in which 2.2 was used.

Table 2.3 is closely related to the sort of empirical tables whose construction has just been discussed. However, this table permits us to make an inference as to the percentage of the population possessing a characteristic, without having an hypothesis about that percentage. Sampling variation may have presented us with an unusual sample in which our inference about the population percentage will be wrong, but if statements are made to the effect that the population percentage lies between the pair of numbers given in the body of the table opposite Number Observed, then only 1 out of 20 such statements should be in error when the 95% Confidence Interval table is used and only 1 out of a hundred when the 99% Confidence Interval table is used. It is to be noticed that additional insurance of making a correct statement has its cost in the length of the interval; the surer the experimenter wishes to be, the larger is the interval and thus the greater is the cost. This table has been prepared on the basis of mathematics and not by sampling from known populations, an empirical method. In the next chapter, some elementary probability is introduced to show how such tables may be prepared and used.

Tests of hypotheses and inferences. In the previous section, two closely related procedures have been outlined. First, the testing of hypotheses was discussed. From a distribution giving the relative frequencies of certain sample results to be expected when a stated hypothesis is true, it is possible to observe which samples must be considered unusual; if the experimenter obtains an apparently unusual sample, he rejects the hypothesis. The percentage of times he rejects the hypothesis when it is true is fixed in advance by the experimenter.

Secondly, an uncertain inference was made. It was not known whether or not the inference was correct, but in the long run only about one in a fixed number, such as 20 or 100, of such statements would be incorrect.

Alternative hypotheses. The fact that one wishes to test an hypothesis is clear admission that the hypothesis may not be correct. In other words, there is an alternative hypothesis or there are alternative hypotheses. Alternatives may be very precise or quite vague. A geneticist may wish to test a 1:1 ratio against the single alternative that the ratio is 3:1; a metallurgist may wish to test the hypothesis that the melting point for an element is the published value against the full set of alternatives involved in the statement that it is other than the published value.

It now appears, then, that it is possible to claim that an hypothesis is true when an alternative hypothesis is true. This is a second type of error, and the experimenter will wish to guard against it. As with the hypothesis, there is some risk which the experimenter must be prepared to take in claiming the hypothesis is true when the alternative is, in reality, the correct one. If the first kind of error discussed is fixed prior to the conduct of an experiment of fixed size, then no control is possible over the second kind of error. However, it is possible to calculate the frequency with which it occurs for various alternative hypotheses. Control over this type of error must be made by the choice of sample or experiment size. This must, of course, be determined prior to the conduct of an experiment where the size is fixed in advance of its conduct.

The hypothesis and its alternative are usually referred to as the Null Hypothesis and the Alternative Hypothesis; the two types of error are customarily called errors of the first kind and errors of the second kind, or Type I errors and Type II errors.

The fact that sample size affects an inference is obvious from Table 2.3. Comsider samples of size 10, 15, 20, 30, 50, and 100 with 60% of the sample observations possessing a specifeed characteristic. One infers that the population percentage lies within the range 26 - 88, 32 - 81, 40 - 77, 45 - 73, and 50 - 70 respectively, the range decreasing with increasing sample size. The fact that sample size affects our tests of hypotheses can be seen from the results of random sampling given in Table 2.2. The range of percentages which we must regard as failing to deny the null hypothesis

2.12 Standard deviation of means. Sample means are variables and, as such, possess means and variances. Either as a matter of observation or intuition, one expects means to be less variable than individuals. Also, if one takes two series of means each based upon a different number of variates, say 10 and 20, it will be found that the variation among the means based upon 20 observations will be less than that based upon 10 observations. Fortunately, as will be demonstrated in Chapter 4, there is a known relationship between s^2 or s among individuals and s^2 or s among means of individuals. The relationship between the standard deviation of a sample mean and the standard deviation of an individual is $s_{\bar{x}} = \frac{s}{\sqrt{n}}$, regardless of the distribution involved. The corresponding relationship between the variances is $s_{\bar{x}} = \frac{s^2}{n}$. Note that $\bar{x}$ is used as a subscript for clarity. Thus the standard deviation of a mean, or, as it is often called, the standard error of the mean, is inversely proportional to the square root of the number of variates entering into the mean. The standard error of the mean for our sample of four variates is $s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{3.37}{\sqrt{4}} = 1.68$. As a matter of computational convenience, $s_{\bar{x}}$ is usually calculated as $\sqrt{s^2/n} = \sqrt{34/3 \times 4}$ $= \sqrt{2.8333} = 1.68$. This estimate, s^2/n, is of the magnitude to be expected of a variance calculated using means of samples of size 4 as the variable. Note that we have only a single mean. Standard deviations are in the same unit of measurement as the observations; e.g., if X_i is a weight in pounds, then s and $s_{\bar{x}}$ are numbers of pounds.

2.13. The linear model. A common model in statistics states that an observation consists of a mean plus an error. The mean may involve a single parameter or be composed of a sum of known or observable multiples of parameters. Further assumptions about the parameters and the errors depend upon the problem at hand. A minimum assumption is that the errors are obtained at random.

Such a model is applicable to the problem of estimating

means and variances as has been discussed so far. We write:

$$X_i = \mu + \epsilon_i .$$

Every observation is an observation on the mean, μ, (Greek letter μ), but is subject to a sampling error denoted by ϵ_i in the case of the i-th observation. For the time being we shall assume that the epsilons (the ϵ_i's) are normally and independently distributed with mean zero and a common variance, σ^2. The normal distribution will be discussed in the next chapter. Independence of the sampling errors is obtained, in this case, by obtaining the sample in a random manner.

The sample mean is

$$\bar{x} = \sum_{i=1}^{n} X_i/n = \frac{n\mu}{n} + \frac{\sum_{i=1}^{n} \epsilon_i}{n} .$$

Clearly, the larger the sample, the smaller is the mean of the epsilons if they are obtained at random, since positive and negative epsilons will tend to cancel each other. This is the same idea that says to expect the variance of means to be smaller than that of individuals, and to expect the means of large samples to be less variable than means of smaller samples. This, in turn, says that the sample mean is a good estimate of the population mean, and that means of large samples are to be preferred over means of small ones.

It is also seen that the linear model implies the possibility of obtaining an estimate of the variance of the epsilons, the real individuals. A finite sample of ϵ_i's would lead to only an estimate of σ^2. Our sample does not permit us to obtain the ϵ_i's since we do not know μ. These, then, must also be estimated, say by corresponding ϵ_i's, calculated as $(X_i-\bar{x})$'s. Some average of these gives an estimate of σ^2. It is usual to use the divisor n-1 because the average of such estimates is σ^2. The use of the calculable $(X_i-\bar{x})$ in preference to a deviation from some other number, in place of the unknown $(X_i-\mu)$, has the advantage that the sum of squares is known to be a minimum. It is sometimes stated that $\bar{x}$ is a <u>minimum sum of squares</u> estimate.

The variability of the epsilons cannot be controlled so we obtain a random sample of epsilons. This random sample supplies us with a valid estimate of the variance which, in the form of a standard deviation, becomes a unit of measurement which can be put to useful work. This work consists of help in deciding when the numerical magnitude of a departure from an hypothesis can be regarded as reasonably due to random sampling. A numerical value, a probability, is used to measure reasonableness. This idea will be given some elaboration in the next section and throughout the book.

2.14 The confidence inference. From the sample data, it is possible to make an inference about the location of the population mean, μ. If a stated degree of confidence is to be associated with the inference, the estimate of μ must be in the form of an interval within which μ is stated to lie. In order to do this, we make use of information to be discussed in Chapters 3 and 4. The sample quantity t is defined there as

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} ,$$

with μ being the only unknown involved. t may be positive or negative. Table 3.8, Snedecor, containes values of t associated with a theoretical t-distribution. For example, opposite d.f. 17 and under .05 is the number 2.110; the table heading tells us that larger sample values of t, either positive or negative, should occur with probability .05 if random samples of size 18 are obtained from a normal distribution. This may be written symbolically as

$$P\ (-2.110 < t < +2.110) = 1 - .05 = .95$$

or

$$P\ (t < -2.110,\ t > +2.110) = .05$$

where P stands for probability, $>$ says that the preceding number (or quantity) exceeds the following, and $<$ says the converse.

Substitution of the definition of t and some algebraic manipulation lead to the symbolic statement:

$$P\left[\bar{x} - t(.05,d.f.)s_{\bar{x}} < \mu < \bar{x} + t(.05,d.f.)s_{\bar{x}}\right] = .95$$

where $t(.05,d.f.) = 2.110$ for sample size 18. This states that μ lies within a certain interval with a probability of .95. Once the sample is obtained and numerical values of $\bar{x}$ and $s_{\bar{x}}$ substituted in the above expression, it becomes a false one, since μ either does or does not lie within the stated interval. This sort of thing is equivalent to flipping a coin, observing the side that is showing, say a head, and stating that the probability of that side being a head is .5. Of course, probability is not involved since the side showing is certainly a head. There was a probability of .5 involved before the coin was flipped but not after the result was observed.

Similarly, there was a probability of .95 that the population mean would lie in the interval to be obtained after the sample was drawn and the necessary calculations performed. Thus, it becomes quite correct to say that if we sample randomly from a normal population, calculate confidence intervals as above, and claim that the interval obtained in each case contains the population mean, then we expect about 5% of our statements to be incorrect. This, in turn, can be restated in each case to be meaningful by saying that the calculated interval contains the population mean unless an unusual sample was obtained, unusual enough to be obtained only about 5 in 100 or 1 in 20 times on the average. 'On the average' is used to make it clear that we don't expect precisely one in each set of 20 statements to be incorrect.

For our sample of four variates $\bar{x} = 7$, $s_{\bar{x}} = 1.68$ and t_{05} for 3df = 3.182 hence $l_1 = 7 - (3.168)(1.68) = 6.5$

$$l_2 = 7 + (3.168)(1.68) = 7.5$$

where l_1 and l_2 denote the lower and upper limits, respectively, of the confidence interval. We say that in the sampled population the mean lies in the interval 6.5 to 7.5, unless a one in twenty mischance in sampling has occurred.

2.15 Coefficient of Variability. The coefficient of variability is the standard deviation expressed as a percentage of the sample

mean, i.e., C.V. = $\frac{100s}{\bar{x}}$. It is a relative measure of variation in contrast to the standard deviation which measures variation in the same units of measure as for the observation. Its usefulness lies largely in the fact that $\bar{x}$ and s tend to vary together, that is, as $\bar{x}$ increases, s does also. From a mathematical point of view, this is unfortunate since it is a clear indication that a normal distribution does not underlie the data. It is desirable to calculate the C.V. for each experiment as it affords a means by which a person can compare the variation found from experiment to experiment. Also since the C.V. is the ratio of 2 averages, it is independent of the unit of measurement used. Thus it is the same whether pounds or bushels are used to measure weight.

It is also useful in estimating the sample size required for a specified degree of precision in a result. This will be illustrated later.

2.16 An Example. We will now illustrate the calculation and the interpretation of the statistics discussed so far by using an actual example. The data in Table 2.4 give the malt extract values on malts made from Kindred Barley grown at 14 locations in the Mississippi Valley Barley Nursery during 1948. Our population can be considered as the malt extract values for the malts made from Kindred barley grown by farmers during 1948 in the area covered by the Mississippi Valley Nurseries. The original values have been modified slightly in order to facilitate certain calculations.

Our first step in the calculation of the arithmetic mean, $\bar{x}$, and the standard deviation, s, using the direct or machine method, is to obtain $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i^2$. These two values can be obtained simultaneously on any of the common types of calculating machines. The arithmetic mean is given by $\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1064}{14} = 76$, and the mean square or variance by

$$s^2 = \frac{\text{sum of squares}}{\text{degrees of freedom}} = \frac{\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2/n}{n-1}$$

$$= \frac{80881.38 - (1064)^2/14}{13} = \frac{17.37}{13} = 1.337.$$

Table 2.4 Malt extract values on malts made from Kindred barley grown at 14 locations in the Mississippi Valley Barley Nurseries during 1948.

Malt extract values		Deviation from mean	Deviates squared
Original	X	$x = x-\bar{x}$	x^2
77.7	77.7	1.7	2.89
76.0	76.0	0	0
76.9	76.9	0.9	0.81
74.6	74.6	-1.4	1.96
74.7	74.7	-1.3	1.69
76.5	76.5	0.5	0.25
74.2	75.0	-1.0	1.00
75.4	75.4	-0.6	0.36
76.0	76.0	0	0
76.0	76.0	0	0
73.9	73.9	-2.1	4.41
77.4	77.4	1.4	1.96
76.6	76.6	0.6	0.36
77.3	77.3	1.3	1.69
Total 1063.2	$\sum_{i=1}^{n} X = 1064.0$	0	$\sum_{i=1}^{n} x^2 = 17.38$
	$\sum_{i=1}^{n} X^2$ 80881.38		$\bar{x} = 1064/14 = 76.0$
	$s^2 = 17.38/13 = 1.337$		$s = 1.156$
	$s_{\bar{x}}^2 = 1.337/14 = 0.0955$		$s_{\bar{x}} = 0.309$

$$C.V. = \frac{(1.156)(100)}{76.0} = 1.5\%$$

The standard deviation which is the square root of the variance is

$$s = \sqrt{s^2} = \sqrt{1.337} = 1.156.$$

The unit of measurement for s is that of the original data. The calculation of the sums of squares by means of the definition formulae $\sum_{i=1}^{n}(X_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2$ is illustrated in Table 2.4.

In section 2.11 it was stated that the interval $\bar{x} \pm s$ on the average will include approximately two-thirds of the variates. This interval for the malt extract values is 76 ± 1.2, 74.8 to 77.2. An examination of the individual values shows that six values, three smaller - 74.6, 74.7, and 73.9 - and three larger - 77.7, 77.4, and 77.3 - lie outside the range 76 ± 1.2.

The arithmetic mean, $\bar{x} = 76$, and the standard deviation, $s = 1.337$, calculated from our sample, provide us with the best available estimates of the corresponding true and unknown parameters of the population, namely μ and σ.

Since we are primarily interested in the population from which our sample was drawn, our next step is to calculate an interval within which we can state our population mean lies and be reasonably certain that our statement is correct. Our first step is to obtain the standard error of our mean, $s_{\bar{x}}$. This is

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n}} \quad \text{or} \quad \frac{s}{\sqrt{n}}$$

$$= \sqrt{\frac{1.337}{14}} \quad \text{or} \quad \frac{1.156}{\sqrt{14}} \qquad \text{respectively}$$

$$= 0.309.$$

The 95% confidence interval is $\bar{x} \pm t_{.05}s_{\bar{x}} = 76 \pm 2.16^{.67}(0.309)$, i.e., 75.33 to 76.67. Note that the tabular value of t used is that for n-1 degrees of freedom, 13 for this problem. Thus we can say that the true malt extract value of our population lies in the interval 75.33 to 76.67. The statement is correct unless we have an unusual sample, one likely to occur about once in 20

times, i.e., unless a one in twenty mischance has occurred in the sampling.

2.17 The use of coding in the Calculation of statistics. Frequently the calculation of statistics can be facilitated by the use of a process known as coding. Coding is useful when it reduces the work involved and promotes accuracy in calculation. A person contemplating its use should be certain that these advantages are sufficiently great to more than counterbalance the time and possibility of error involved in the process. Coding consists of replacing each variate by a number on a new scale, the scale being the same for all variates, using one or more of the following operations: addition, subtraction, multiplication and division.

The arithmetic mean is affected by every coding operation. for example, if the variates are coded by first multiplying by 10 and then subtracting 100, the mean of the coded numbers must be increased by 100 and then divided by 10. The rule is to apply the inverse operations in the reverse order.

The standard deviaiion is affected only by the operations of multiplication and division. The addition or subtraction of a value to each variate does not affect measures of variation. This is to be expected since it merely shifts the origin of the observations without affecting their spread. Since both multiplication and division change the unit of measurement, measures of variation calculated from coded variates are decoded by applying the inverse operation to the coded numbers.

We will illustrate the use of coding in the calculation of the the arithmetic mean and the standard deviation for the malt extract values given in Table 2.4. In this example the coded values will be obtained by subtracting 70 from each variate, then multiplying each by 10. Thus the first variate $X_1 = 77.7$ is replaced by its coded value, $X'_1 = (77.7-70)10 = 77$. The arithmetic mean and the standard deviation of the coded numbers are then obtained using the procedures illustrated in 2.16. $\sum_{i=1}^{n} X'_i = 840$ and $\sum_{i=1}^{n} X'^2_i = 52,138$.

The arithmetic mean of the coded values is $\bar{x}' = \frac{840}{14} = 60$. To

decode $\bar{x}'$, apply the inverse operations in the reverse order, namely divide by 10 and add 70. Thus $\bar{x} = \frac{\bar{x}'}{10} + 70 = \frac{60}{10} + 70 =$ 76, which is the same value we obtained in section 2.16. The standard deviation in coded values is determined as follows:

$$s' = \sqrt{\frac{\sum_{i=1}^{n} X_i'^2 - (\sum_{i=1}^{n} X_i')^2/n}{n-1}} = \sqrt{\frac{52,138 - (840)^2/14}{13}} = 11.56.$$

To decode s', we need only to divide s' by 10. Thus $s = \frac{s'}{10} =$ $\frac{11.56}{10} = 1.156$, the value obtained in section 2.16. The accuracy of the calculation is in no way affected by the coding process used here.

2.18 The Frequency Table. Where the sample consists of a relatively large number of observations it is usually desirable to summarize the data in the form of a frequency table. This is a table which shows the frequencies with which the variates fall into certain classes. It is of value in that it reduces the mass of raw data into a more digestible form and also provides the basis for graphical presentation of the data. Statistics such as the mean and standard deviation can be calculated with much less work using a frequency distribution than from the original values. This is particularly true if an electric calculator is not available.

Both continuous and discontinuous data can be summarized in a frequency table. In the case of discontinuous variables such as the number of heads recorded in the toss of five coins, the class values to be used are generally obvious. Thus the frequency table for the number of heads occurring in the tossing of 5 coins 100 times could be given as:

Class values	Frequency
5	4
4	15
3	29
2	30
1	17
0	5
Total	100

In cases where the number of possible classes is large it is often desirable to reduce their number. For example, if 20 coins were tossed in place of 5 and the number of heads recorded, we could use class values of 0-2 heads, 3-5 heads, etc.

With continuous variates the classes have to be chosen in some arbitrary manner. The choice will depend upon a number of factors such as number of observations, the range of variation, the accuracy required in the calculation of statistics from the table, and the degree of summarization of the data necessary so that small irregularities will not cover up general trends. The last two points usually work in opposite directions. The greater the number of classes the greater the accuracy of any calculations made from the table; however, if the number of classes is too great the data will not be summarized sufficiently. Thus in deciding upon the size of the class interval, a balance must be reached between these two opposing factors.

A rule that is often used in determining the size of the class interval where extreme accuracy is required in calculations made from the resulting frequency table is to make the interval not greater than one-quarter of the standard deviation. If this rule is strictly adhered to, the data is often not sufficiently summarized for graphical presentation. If the size of the class interval is increased to one-third to one-half of a standard deviation, the resulting frequency table will usually be a sufficient summary for graphical presentation and for most sets of data. The error introduced into any statistics calculated from the table will be so small that it can be ignored. Since the standard deviation is not known at the time a frequency table is being prepared it is necessary to estimate it. Tippett ( ) published in Biometrika detailed tables showing the relationship between the range and the standard deviation in samples from normal populations. This has been condensed in a short table by Goulden ( ) and is given below is Table 2.5.

Table 2.5

Values of the ratio, range divided by the standard deviation(s), for sample sizes from 20 to 1000

Number in sample	Range/s	Number in sample	Range/s
20	3.7	200	5.5
30	4.1	300	5.8
50	4.5	400	5.9
75	4.8	500	6.1
100	5.0	700	6.3
150	5.3	1000	6.5

See Sturgis, H.A. JASA Mar 1936 = ? = $\frac{\text{range}}{1 + 3.322 \log 10N}$

To illustrate the preparation and use of a frequency table, we will use the yield to the nearest gram of 229 spaced plants of Richland soybeans reported by Drapala ( ). The plants ranged in yield from 1 to 69 grams. From Table 2.5 we find that the ratio $\frac{\text{range}}{s}$ for 229 is about 5.6. Thus the estimated s would be $s = \frac{68}{5.6} = 12.2$. One-third of the estimated s is 4 and one-half is 6. Since 12.2 is only an approximation of s the calculated class interval can be adjusted somewhat to make it more usable. It is convenient to make the class interval an odd number rather than an even, as the mid-point of such an interval requires one less decimal point. We will select 5 as our class interval. If we had been interested in greater accuracy in the calculation of the statistics from the table we would have used an interval of 3 units which is about 1/4 of the estimated s. In setting up the class values it is desirable to make the lower limit of the first class slightly less than the smallest value in the sample.

In the calculation of certain statistics from frequency tables, Sheppard's corrections for grouping are sometimes used. Considered as average adjustments, they remove bias from certain statistics that have been calculated from a frequency table in place of the individual values. For example, in the calculation of s^2 from frequency tables, the bias is positive and is equal to

1/12 of the class interval. Usually these corrections are ignored. They are generally small but may adjust in the wrong direction if the rounding lattice is not imposed at random.

Sheppard's corrections are for removing a definite bias and in no way do they make allowance for inaccuracies due to errors in grouping. Errors due to grouping can be seen if we consider the 5 variates which enter into the second class to Table 2.6. These values are 6, 7, 7, 9, and 9. The arithmetic mean of these values is 7.6, while in the calculations using Table 2.6 they are given a value of 8. If we examined each class for this type of error, we would find that about half would have a negative error and half a positive error and thus they tend to cancel each other.

After the class interval has been chosen, the next step is to set up the necessary classes and sort out the variates accordingly. Two methods of sorting out the variates are commonly used. The first is referred to as the tally score method which is illustrated below using the first few classes of Table 2.6.

Class range	Class value or midpoint	Tally score	Frequency
1 - 5	3	𝍸 //	7
6 - 10	8	𝍸	5
11 - 15	13	𝍸 //	7

This method consists of making a stroke in the proper class for each variate and then summing these for each class to obtain the frequency. It is customary to place each fifth stroke through the preceding four, as shown. This is a convenience in counting. The method has the disadvantage that errors discovered upon checking are difficult to find.

A sounder method is to write the value of each variate on a card of a size convenient for handling. The class ranges are then written out on other cards and arranged in order on a desk. The variate cards are then sorted out into their respective classes. Checking can be accomplished readily by examining the cards in each of the classes. It is very important that the values be

entered on the cards accurately. The frequency in each class can be determined by counting the cards in each of the classes. If IBM equipment is available the value of each variate can be punched on a card and these cards can be readily sorted by an IBM sorter.

Table 2.6 is an example of a frequency table showing the yield in grams of 229 spaced plants of Richland soybeans reported by Drapala ( ).

Table 2.6

Frequency table for yield in grams of 229 spaced plants of Richland soybeans

Yield in grams	3	8	13	18	23	28	33	38	43	48	53	58	63	68
Frequency	7	5	7	18	32	41	37	25	22	19	6	6	3	1

2.19 Calculation of the mean and standard deviation from a frequency table. These calculations will be illustrated by the frequency distribution given in Table 2.6. Coding is included. The first step is to prepare a table such as Table 2.7. The first column X consists of actual class values. The second column X' is formed by replacing actual values by coded values. To facilitate the work, zero is assigned to the middle class value. The other class values increase by unity in one direction and decrease by unity in the other direction. Column B is the frequency and the last two columns are obtained as explained by their headings.

The sums of the last three columns, Σf_i, $\Sigma f_i X'_i$, and $\Sigma f_i X'^2_i$ respectively, are the values necessary for the calculation of the mean and the standard deviation. They correspond to n, $\sum_i X_i$, and $\sum_i X_i^2$ respectively. Note that i goes from 1 to 14 for the frequency table and from 1 to 229 for the data. We shall no longer be specific about this point unless it is not clear by the context; the subscript will also be abandoned at times.

The arithmetic mean, $\bar{x}$, and the standard deviation, s, expressed in actual units of measurement can be calculated from the coded values by the following formula:

$$\bar{x} = a + I\frac{(fX')}{n} = 33 + 5\frac{(-49)}{229} = 31.93$$

$$s = I\sqrt{[\Sigma fX'^2 - (\Sigma fX')^2/n]/(n-1)} = 5\sqrt{6.57} = 12.80$$

where I is the class interval and a is the assumed mean, namely the X_i value corresponding to the coded class value of 0.

Table 2.7

Class value or midpoint of class range		Frequency	Frequency multiplied by coded class value	Frequency multiplied by square of coded class value
Actual	Coded			
X_i	X'_i	f_i	$f_i X'_i$	$f_i X'^2_i$
3	-6	7	-42	252
8	-5	5	-25	125
13	-4	7	-28	112
18	-3	18	-54	162
23	-2	32	-64	128
28	-1	41	-41	41
33	0	37	0	0
38	+1	25	25	25
43	+2	22	44	88
48	+3	19	57	171
53	+4	6	24	96
58	+5	6	30	150
63	+6	3	18	108
68	+7	1	7	49

$\Sigma f = 229 \quad \Sigma fX' = -49 \quad \Sigma fX'^2 = 1507$

$$\bar{x} = a + I\frac{(fX')}{n}$$

$$= 33 + 5\,\frac{(-49)}{229}$$

$$= 31.93$$

$$s'^2 = [\Sigma fX'^2 - (\Sigma fX')^2/n] \,/\, (n-1)$$

$$= [1507 - (-49)^2/229] \,/228$$

$$= 6.57$$

$$s = I\sqrt{s'^2} = 5\sqrt{6.57} = 12.80$$

2.20 Graphical representation of a frequency distribution. Two types of graphs are in general used to represent data from a frequency table. The best and the one most commonly used is the histogram. It is a picture of a frequency table in which the class values are represented along the horizontal axis, and a rectangle above the class interval represents a frequency. The histogram is valuable in that it presents data in a form that most people readily understand and in which they see at a glance the general nature of the distribution. If it is desired to compare an actual distribution with a theoretical distribution, the theoretical curve can be superimposed on the histogram and any discrepancies can be readily ascertained.

The frequency polygon is prepared by locating the midpoint of each class value and marking a point above it at a height determined by the frequency. These points are then connected by straight lines. The frequency polygon does not give as accurate a picture for the sample as does the histogram, but tends in its shape to give the smooth curve of the population from which the sample was drawn. The histogram and frequency polygon for the data given in Table 2.7 are shown in Figure 2.1.

It is important in the preparation of both the histogram and frequency polygon that the number of classes be sufficiently large that the general shape of the distribution can be readily ascertained, yet not have too many classes so that too much detail is given. For most data, this will be accomplished by using a class interval which is one-third to one-quarter of a standard deviation. The number of classes should be between 8 and 20.

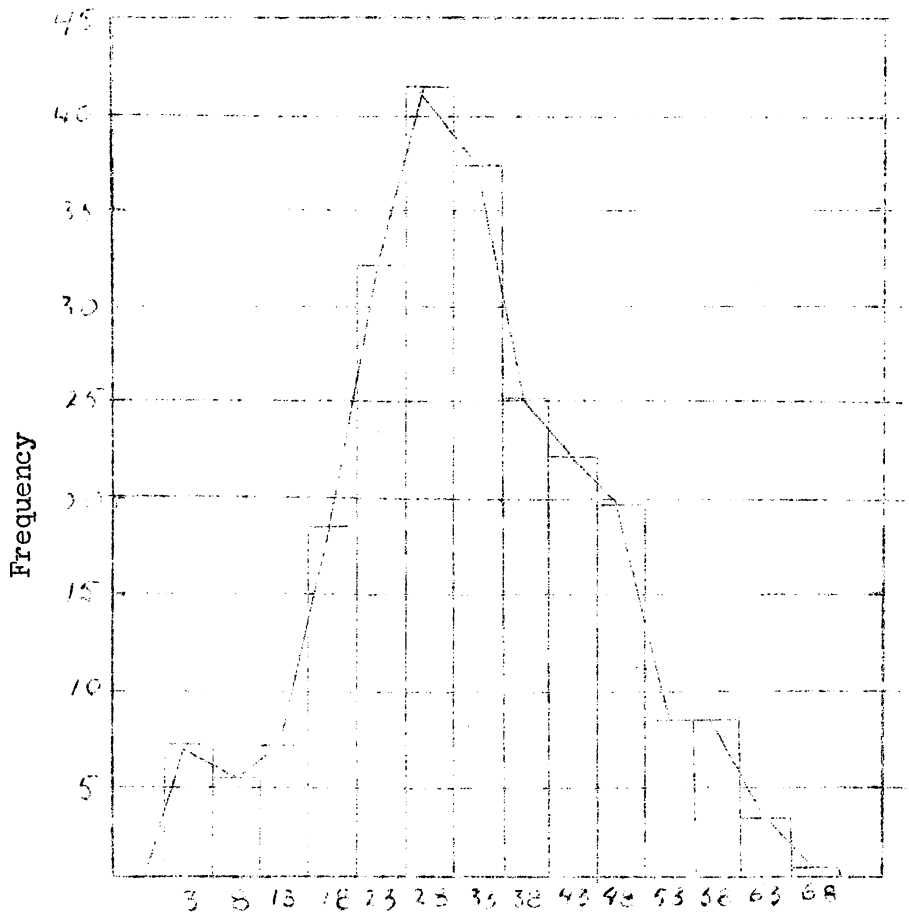


Figure 2.1 - Histogram and Frequency Polygon for the data of Table 4

CHAPTER IV

SAMPLING FROM A NORMAL DISTRIBUTION

4.1 Introduction. In chapter 2 the calculation of statistics of central tendency and dispersion for sample data was discussed. For the procedures given, it was stressed that the statistics, when calculated from a random sample from a single population, provide unbiased estimates of the corresponding population parameters. It was stated that the sample standard error of the mean can be obtained from the sample standard deviation by means of a known relationship. A method for using the sample data to establish a range about the sample mean, known as the confidence interval, within which the population mean is stated to lie, was given. The average percentage of incorrect statements can be fixed in advance and is associated with a probability level. The method uses the mean, the standard error of the mean, and a so-called t-value.

In chapter 3 some theory relating to both the normal and t distributions was given. We are now ready to demonstrate by sampling methods the appropriateness of certain of the relationships that were used in previous chapters. This sort of approach is referred to as the empirical method.

In this chapter we examine the distribution of the following statistics: $\bar{x}$, s^2, s, $s_{\bar{x}}$, $s_{\bar{d}}$, and t; show that these statistics give unbiased estimates of the population parameters, and demonstrate the interrelationships between the statistics of dispersion. This is done by using data from random samples drawn from a normally distributed population.

By this sampling approach it is possible to verify many of the important theorems and principles involving probability which have been developed by mathematical statisticians. An understanding of these theorems and principles is necessary in order to appreciate the material discussed in later chapters.

4.2 A normally distributed population. Table 4.1 consists of the yields in pounds of butter fat produced by one hundred Holstein cattle. The original data have been modified somewhat to form an approximately normal distribution. In chapter 3 it was stated that a normal population has a continuous variable and is, consequently, not finite in size, and that the range is infinite. In other words, a variate can assume any value. The data in Table 4.1 depart from

Figure 4.1.a
Graphical representation of the array of pounds of butterfat of 100 Holstein cows

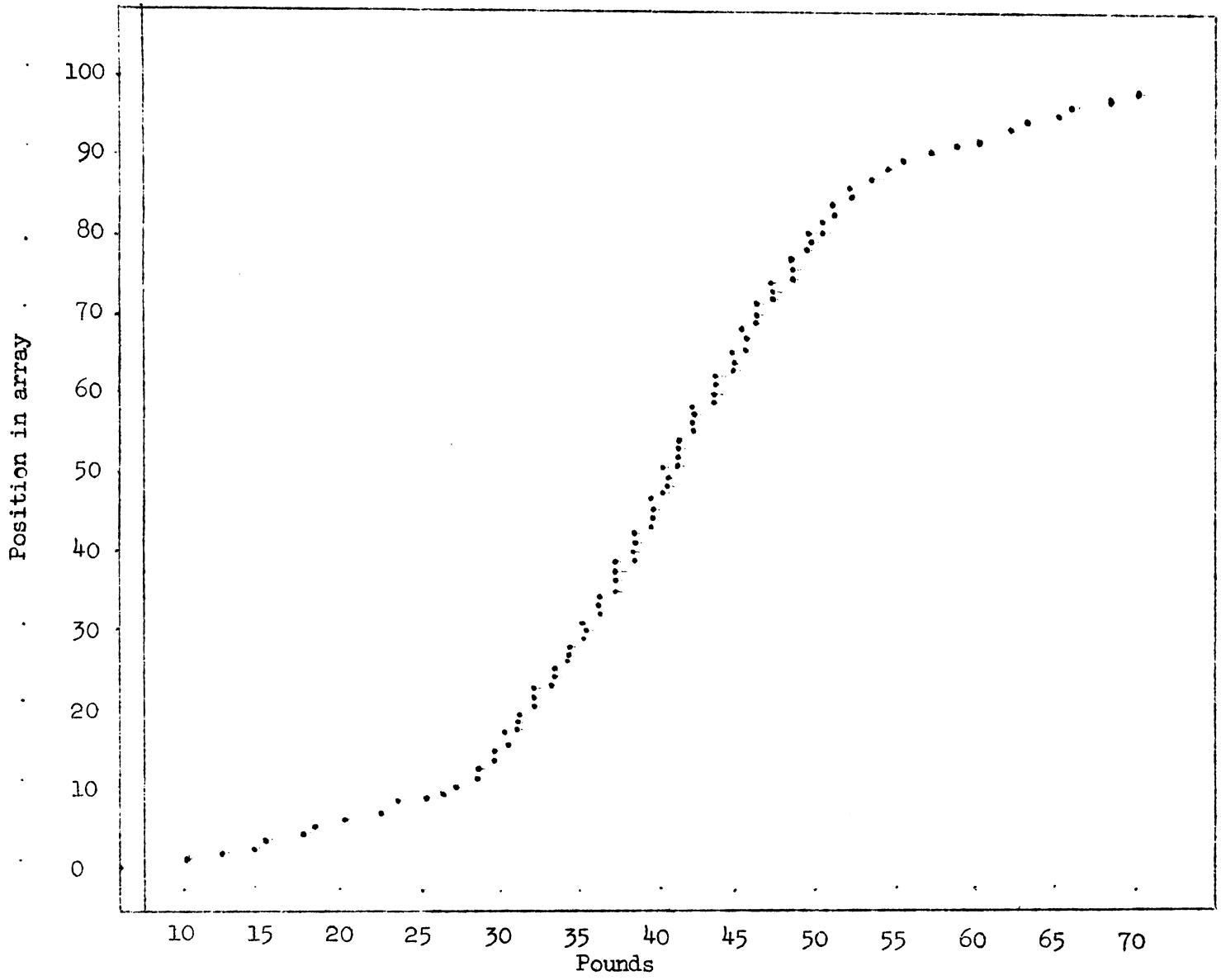
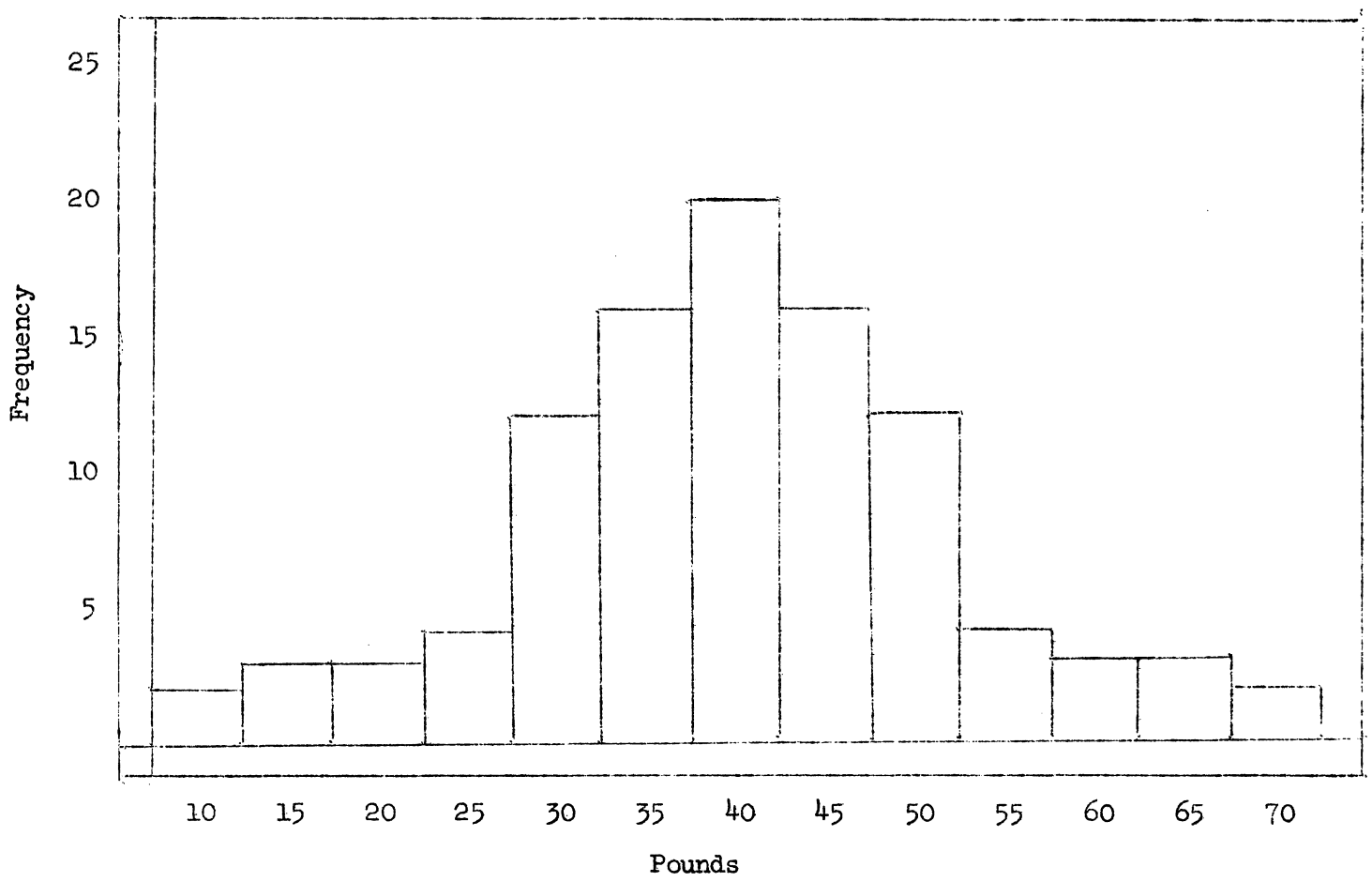


Figure 4.1.b

Histogram of the distribution of pounds of butterfat from 100 Holstein cows



the model in two major points: the variates have a finite range and are discontinuous. The fact that there are a finite number of values is not serious since we can return any value to the population before drawing the succeeding one. This is equivalent to dealing with an infinite population. The effects due to the finite range and discreteness of the data are small in comparison with sampling variation and accordingly will have little or no effect upon our conclusions. The salient characteristics of the distribution are depicted in Figure 4.1, a and b. Figure 4.1.b is a frequency histogram of the data with pounds of butterfat on the horizontal axis and the frequency on the vertical axis. The values show a concentration at the center and thin out symmetrically on both sides, slowly at first and then more rapidly. Figure 4.1.a shows the 100 values in the form of an array. The relation of the histogram to the array is that the height of the rectangle in any class of the histogram is proportional to the number of dots lying within the vertical lines of the array.

Table 4.2 is a frequency distribution of pounds of butterfat for the data of Table 1. Each class has a range of five pounds.

Table 4.2

Frequency distribution of pounds of butterfat of 100 cows

Midpoint class mark	10	15	20	25	30	35	40	45	50	55	60	65	70	
Frequency	2	3	3	4	12	16	20	16	12	4	3	3	2	100

4.3 Random samples from a normal distribution. The 100 items in Table 4.1 have been assigned the numbers 00 to 99 in order to facilitate drawing random samples by use of the table of random numbers given in Table 2.1. The discussions which follow are based upon 500 samples of 10 observations each drawn at random from Table 4.1.

A suitable procedure for drawing random samples from a table such as Table 4.1 is as follows: Place your pencil on some digit in the table of random numbers, Table 2.1, then use this and the next three digits to determine the intersection of a row and column in the same table. For example, suppose that the four digits selected were 2 3 7 1, then start in row 23 at column 71 and move in any direction,

recording the integers in pairs. Twenty numbers give 10 pairs which are to serve as the item numbers in Table 4.1. If you move downwards in Table 2.1, the integers recorded in pairs will be 79, 04, 21, 65, etc. The pounds of butterfat in Table 1 corresponding to these random integers, namely 49, 17, 32, 44, etc., are now recorded. To arrive at a starting point for the next sample, simply use the last two pairs of integers, 67 and 21, to locate the row and column.

This procedure assures that each item may be drawn any number of times. The result should be equivalent to drawing from a bag of 100 beans marked with the hundred butterfat yields, each bean being replaced in the bag and the beans being thoroughly mixed before making the next draw. Thus, sampling is always from the same population and the probability of drawing any particular item is practically the same. Either procedure gives essentially the same results as if the drawings were made from an infinite population.

It is suggested that each student draw at least ten samples in the above manner, and that the data obtained by all students in the class be pooled in order to provide sufficient data to illustrate the material discussed in this chapter. Table 4.3 gives 5 such samples together with certain calculations discussed in this chapter.

4.4 The distribution of sample means. Five hundred samples of 10 observations were drawn from Table 4.1. The frequency distribution of the means of these samples is given in Table 4.4. A class interval of 1 1/2 pounds was used. This distribution illustrates several basic features of sampling. First, the distribution of the means is approximately normal, as may be seen by comparing the observed with the theoretical frequency. The theory is that the derived distribution of sample means obtained from observations drawn from a normal population is likewise normal. It is also true that the distribution of means tends to be normal even if the distribution of the variates is considerably anormal. This is very important in practice for the form of the parent distribution is rarely known. A second feature is that the range of the means is considerably less than that of the individuals, being 27 pounds as compared to 60 pounds for the individual variates, i.e.means are less variable than are individuals. Thirdly, the average of the means, 39.79 pounds, is very close to 40 pounds, the true population mean. This illustrates clearly the nature of unbiasedness. The sample mean is said to be unbiased because the mean of all possible

sample means is the population mean. For 500 sample means, we find a mean of means of 39.79 pounds, very close to the population mean, $\mu = 40$ pounds.

4.5 The distribution of sample variances and standard deviations. For each of the 500 samples of 10, the variance, s^2, and the standard deviation, s, were calculated. The procedure and results are illustrated for five such samples in Table 4.3.

The distribution of the 500 sample variances is given in Table 4.5. It is skewed with a clustering to the left of the mean of the 500 s^2's, denoted by $\overline{s^2}$, and has a long tail to the right. This distribution is similar to that of χ^2 (chi square) which will be discussed later. The quantities $\frac{(n-1)s^2}{\sigma^2} = \frac{9s^2}{144}$ for our samples, are distributed as χ^2 with $n-1 = 9$ d.f. In spite of the skewness of the distribution, the mean variance is 140.4, closely approximating the population variance, $\sigma^2 = 144$. This demonstrates the unbiasedness of s^2 as an estimate of the population variance, σ^2. Notice the variability among the individual s^2's which exhibit a range from 20 to 380.

Table 4.5

Frequency distribution of 500 variances (s^2) for samples of size 10 drawn at random from the butterfat data given in Table 4.1

Class mark	20	40	60	80	100	120	140	160	180	200	220	240	260	280	300	320	340	360	380
Frequency	11	27	40	46	59	62	55	51	43	27	21	16	18	7	5	7	3	1	1

$\overline{s^2} = 140.4$, using d.f. $\sigma^2 = 144$

$(= 126.4$, using sample size$)$

Table 4.6 gives the distribution of the observed standard deviations, s. It shows a slight skewness, less than that of the variances. The reason for the greater skewness in the distribution of s^2 as compared to s is that the variances above the mean increase disproportionately to those below the mean, when obtained as the squares of the standard deviations. This can readily be seen by an examination of the values given below.

s	s^2
10	100
11	121
12	144
13	169
14	196

Table 4.3

Five samples of 10 observations drawn at random from the butterfat data of Table 4.1, together with sample statistics

Sample item number and formulas	Sample number: Item number and sample value									
	1		2		3		4		5	
	80	49	39	37	48	40	53	41	70	46
1	96	65	39	37	95	63	39	37	94	62
2	37	37	51	40	59	42	63	43	42	38
3	04	17	34	36	54	41	84	51	16	30
4	29	34	81	49	47	39	16	30	08	23
5	84	51	34	36	41	38	17	30	59	42
6	05	18	49	40	81	49	67	45	97	66
7	71	46	47	39	09	25	98	68	98	68
8	35	36	75	47	03	15	97	66	45	39
9	03	15	23	32	15	29	59	42	89	54
10	69	45	96	65	87	52	00	10	50	40

Sum	= ΣX	364	421	393	422	462
Mean	= $\bar{x}$	36.4	42.1	39.3	42.2	46.2
	ΣX^2	15626.00	18541.00	18615.00	22009.00	23498.00
C.F.	= $(\Sigma X)^2/10$	13249.60	17724.10	15444.90	17808.40	21344.40
S.S.	= $\Sigma X^2-(\Sigma X)^2/10$	2376.40	816.90	3170.10	4200.60	2153.60
s^2	= S.S./9	264.04	90.77	352.23	466.73	239.29
s	= $\sqrt{s^2}$	16.2	9.5	18.7	21.6	15.4
$s_{\bar{x}}$	= $\sqrt{s^2/10}$	5.03	9.52	5.93	6.83	4.89
t	= $(40-\bar{x})/s_{\bar{x}}$	0.72	0.22	0.12	0.32	1.27
$t_{.05}s_{\bar{x}}$	= $2.262\ s_{\bar{x}}$	11.38	21.53	13.41	15.45	11.06
C.L.	= $\bar{x} \pm t_{.05}s_{\bar{x}}$	25.02 - 47.78	20.57 - 63.63	26.89 - 52.71	26.75 - 57.63	35.14 - 57.26

The differences between s's are one unit whereas the differences between successive s^2's are 21, 23, 25, and 27, consecutive odd numbers.

Table 4.6

Frequency distribution of 500 standard deviations (s) corresponding to the variances of Table 4.5

Class mark	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Frequency	1	10	14	23	37	42	55	69	66	63	40	26	30	11	11	2

$\bar{s} = 11.47$ / $s^2 = 11.85$ $s = 12$

The average of the 500 standard deviations is 11.47 as compared to $\sigma = 12$. The square root of the average of the 500 variances is $\sqrt{140.4} = 11.85$. The difference between the two values is due to a small bias in s which results in an underestimate of σ. Consequently, if we take an average of many s's, we can be quite confident that the result will be less than the population value being estimated.

In individual samples, the bias is negligible and can be ignored. However, if it is desired to obtain an average of several standard deviations, a better estimate of s is obtained by extracting the square root of the average of the variances rather than by averaging the standard deviations.

4.6 An illustration concerning degrees of freedom. In chapter 2 it was stated that the calculation of s^2 from a sample results in an unbiased estimate of σ^2 when the divisor of the sum of squares is the degrees of freedom, n-1, rather than n. This is so since $\sum_i (X_i - \bar{x})^2$ underestimates $\sum_i (X_i - \mu)^2$ on the average by a factor of $(n-1)/n$. This will now be demonstrated by using the average sum of squares for the 500 samples of 10. The average s^2 for these data is 140.42 when the variance is calculated by the formula, sum of squares divided by degrees of freedom, viz. $s^2 = \frac{\Sigma(X_i - \bar{x})^2}{n-1}$; consequently $\Sigma(X_i - \bar{x})^2 = (n-1)s^2$. Thus, when the degrees of freedom are the same for each sample, the average of the sums of squares is the product of the average variance and the degrees of freedom. Consequently, the average of the sums of squares for our 500 samples of 10 is 140.42 x 9 = 1263.78. If n had been used in place of n-1, the average variance would have been 1263.78/10 = 126.4, a much smaller value than 140.4. The difference between the values obtained using n and n-1, obviously, becomes less as n increases. However, no matter how large n is, the

Table 4.4

Frequency distribution of 500 means of samples of 10 items drawn at random from the butterfat data of Table 4.1

Class mark (pounds)	Frequency	Theoretical frequency
26.5	1	
28	0	
29.5	2	
31	2	
32.5	14	
34	20	
35.5	47	
37	65	
38.5	74	
40	71	
41.5	78	
43	49	
44.5	40	
46	24	
47.5	8	
49	4	
50.5	0	
52	0	
53.5	1	

$\bar{x} = 39.79$

unbiased estimate of σ^2 is found by dividing the sample sums of squares by the degrees of freedom.

4.7 The standard error of standard deviation of the mean. The standard error of the mean, $s_{\bar{x}}$, is one of the most useful statistics available. It is calculated as

$$s_{\bar{x}} = s / \sqrt{n} \quad \text{or} \quad s_{\bar{x}} = \sqrt{s^2/n},$$ i.e. from s or s^2. $s_{\bar{x}}$ provides us with an unbiased estimate of $\sigma_{\bar{x}}$, the variation among sample means of size n drawn with observations from a population with standard deviation σ. Thus for each sample of size 10, $s_{\bar{x}}$ is an estimate of $\sigma_{\bar{x}} = \sigma/\sqrt{10} = 12/\sqrt{10} = 3.79$ pounds. To obtain the most accurate estimate of $\sigma_{\bar{x}}$ from our 500 samples of 10 observations, calculate $s_{\bar{x}}$ by extracting the square root of the average of our variances divided by n= 10. Thus, we have

$$s_{\bar{x}} = \sqrt{\bar{s}^2/n} = \sqrt{140.4/10} = 3.75 \text{ pounds.}$$

This is a better estimate of $\sigma_{\bar{x}}$ than that obtained by dividing the average of the 500 standard deviations, an average of biased estimates, by the square root of 10, i.e. as

$$s_{\bar{x}} = \bar{s} / \sqrt{n} = 11.47 / \sqrt{10} = 3.63 \text{ pounds.}$$

Where a single value of s is available, $s/\sqrt{n}$ and $\sqrt{s^2/n}$ are identical.

The standard errors of the mean as calculated from different samples show sample variation. However, it is seen that the average $s_{\bar{x}} = 3.75$ pounds is a very close estimate of $\sigma_{\bar{x}} = 3.79$ pounds. In order to justify obtaining $s_{\bar{x}}$ from s, we will calculate the standard deviation for our 500 means from samples of size 10 as

$$s_{\bar{x}} = \sqrt{\frac{\Sigma\bar{x}^2 - (\Sigma\bar{x})^2/500}{499}} = 3.71.$$

The close agreement between this and the more accurate of the previous estimates of s, enables us to state with more confidence that there is this definite relationship between σ and $\sigma_{\bar{x}}$, namely that $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ and accordingly that each random sample provides an estimate, $s_{\bar{x}}$, of the standard error of the mean, $\sigma_{\bar{x}}$.

It is important to realize that whereas the variance of a mean decreases inversely by n, that the standard error of a mean decreases inversely as the $\sqrt{n}$. This is clearly shown by

example as well as formula:

n	s_x^2	$s_{\bar{x}}$
4	$\frac{s^2}{4} = \frac{144}{4} = 36$	$\frac{s}{\sqrt{4}} = \frac{12}{\sqrt{4}} = 6$
8	$\frac{s^2}{8} = \frac{144}{8} = 18$	$\frac{s}{\sqrt{8}} = \frac{12}{\sqrt{8}} = 4.24$
16	$\frac{s^2}{16} = \frac{144}{16} = 9$	$\frac{s}{\sqrt{16}} = \frac{12}{\sqrt{16}} = 3$

The importance of this will be seen when we discuss methods used to determine whether or not there are real differences among means associated with different treatments.

4.8 Distribution of t. The use of t and the nature of the t distribution have already been discussed and explained. We are now ready to show that the distribution of our 500 sample t-values approximate the theoretical distribution of t for nine degrees of freedom.

For each of the 500 samples of 10, t was calculated by the formula $t = (\bar{x}-\mu)/(s_{\bar{x}}) = (\bar{x}-40)/(s_{\bar{x}})$. Thus t is an expression for the deviation of the sample mean from the population mean in terms of sample standard deviation units, a natural unit of measurement for making certain decisions about the usualness or unusualness of the deviation. Since our sample maans are distributed fairly symmetrically about μ, approximately one-half of the t-values will be positive and the mean of all the t-values will be approximately zero. For our 500 samples, 248 t-values were positive and 252 negative with a mean value of -0.038.

Table 4.7 is a frequency distribution of the observed t-values. The class intervals, which are unequal, were selected to compare the observed frequencies with the theoretical frequencies of t for d.f. = 9, using available tables. Thus, the class boundaries are identical with those for tabulated t at the 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.02, and 0.01 probability levels given in the t-table. The percentage frequencies for the sample values as well as those for the theoretical values of t are given to facilitate comparison.

In a population of t-values, 2.5% are larger than +2.262 and 2.5% are smaller than -2.262. This is seen from the theoretical percentage frequency. The last column in Table 4.7 combines both

Table 4.7
Sample and theoretical values of t for 9 degrees of freedom

$$t = \frac{\bar{x}-\mu}{s_{\bar{x}}}$$

Interval of t		Sample		Theoretical		
			Percentage	Percentage	Cumulative	
From	To	Frequency	frequency	frequency	One tail	Both tails
----	-3.250	2	0.4	0.5	100.0	
-3.250	-2.821	2	0.4	0.5	99.5	
-2.821	-2.262	7	1.4	1.5	99.0	
-2.262	-1.833	12	2.4	2.5	97.5	
-1.833	-1.383	29	5.8	5.0	95.0	
-1.383	-1.100	21	4.2	5.0	90.0	
-1.100	-0.703	63	12.6	10.0	85.0	
-0.703	0.0	105	21.0	25.0	75.0	
0.0	0.703	144	28.8	25.0	50.0	100.0
0.703	1.100	38	7.6	10.0	25.0	50.0
1.100	1.383	30	6.0	5.0	15.0	30.0
1.383	1.833	23	4.6	5.0	10.0	20.0
1.833	2.262	15	4.0	2.5	5.0	10.0
2.262	2.821	6	1.2	1.5	2.5	5.0
2.821	3.250	1	0.2	0.5	1.0	2.0
3.250	-----	2	0.4	0.5	0.5	1.0
		500	100.0			

tails of the distribution by ignoring the sign of t. This is the column usually referred to when talking about probability levels. Thus 2.262 is referred to as the value of t at the 5% level. When only the positive tail of the t-distribution is considered, 5% of the t's lie beyond 1.833; on the other hand, when both tails are considered, 1% of the t-values lie beyond $\pm$ 3.250, the t-value at the 1% level of t for 9 degrees of freedom. An examination of the sample values, considering both tails, shows that 20 t's exceed the 5% level and 4 t's exceed the 1% level as compared with expected numbers of 25 and 5 respectively. This shows a reasonable agreement between the sample and theoretical values. A comparison of the sample and theoretical values at other levels of probability also shows a close agreement.

4.9 The confidence statement. We are now ready to check on the confidence statements made from our samples. For each sample we establish, for any level of probability, an interval about the sample mean within which we state that the true population mean will fall, the percentage of correct statements depends upon the level of probability chosen. Thus, if we solve $\pm t = (\bar{x}-\mu)/(s_{\bar{x}})$ for μ, we get $\mu = \bar{x} \pm ts_{\bar{x}}$. The two values of μ are denoted by l_1 for the lower limit and l_2 for the upper limit. Thus $l_1 = \bar{x} - ts_{\bar{x}}$ and $l_2 = \bar{x} + ts_{\bar{x}}$. If we desire to establish an interval about each sample mean, state that the population mean lies in the interval, and have 95% of our statements correct on the average, then we use the 5% tabulated value of t for n-1 degrees of freedom. The 1% tabulated value of t is used for similar statements of which 99% will be correct on the average. Since the latter value of t is larger, it is seen that the confidence interval is larger. Thus, the cost of an increasing proportion of correct answers is wider confidence intervals.

For each of the 500 samples of 10, $l_1 = \bar{x} - 2.262s_{\bar{x}}$ and $l_2 = \bar{x} + 2.262\ s_{\bar{x}}$ have been calculated. Since it is known that $\mu = 40$ pounds, the number of correct statements regarding μ can be determined readily. For the 500 samples, 480 gave correct statements at the 5% level and 496 at the 1% level. This compares reasonably well with the theoretical values of 475 and 495 respectively. The percentage of false statements is the same as the percentage of sample t-values which exceed the 5% or 1% tabular t-values.

It is important to remember that the statements made at the 5% level are right unless a one in twenty mischance has occurred in sampling; similarly for the 1% level.

In actual practice the parameter μ is not known. Accordingly, an experimenter never knows whether μ lies in the confidence interval. However, he does know of the percentage of inferences regarding μ that will be correct.

An erroneous idea that is sometimes held is that a 95% confidence limit about a sample mean gives the range within which 95% of future sample means will fall. This is obviously wrong for the distribution of sample means is centered on the population mean and not upon a particular sample mean.

4.10 The sampling of differences. A problem which often confronts an experimenter is that of determining whether there is a difference in response to two treatments. The hypothesis usually set up is known as the null hypothesis and states that there is no difference between the means of the two populations sampled, that is, that the samples can be considered to be drawn from the same population if it is known that the variances are the same. We are now ready to consider sampling from a normal distribution when this null bypothesis is true and there is a single variance. This will be done by sampling from a normally distributed population of differences whose mean is zero. This sampling procedure can be related to the real but special case where observations are obtained as pairs of values, one associated with each treatment. Such a population is readily set up by making use of a theorem which states: If items are drawn at random from a normal population then randomly paired, the differences between the individuals of the pairs are normally distributed with a mean of zero.

Accordingly, the 500 samples of 10 observations drawn at random from Table 4.1 were paired at random and the differences obtained. For each of the resulting 250 samples of 10 differences the mean difference, $\bar{d}$, the variance of the differences, s_d^2, the standard deviation of the differences, s_d, the standard error of the mean difference, $s_{\bar{d}}$, the t-value and confidence limits for the population mean difference were calculated. The procedure used to obtain and treat the differences, d, is illustrated for 3 sets in Table 4.8, which is similar to Table 4.3 for the individual values.

Table 4.8

Three samples of differences from observations drawn at random from Table 4.1

Item Numbers	Paired Observations	Differences	Item Numbers	Paired Observations	Differences	Item Numbers	Paired Observations	Differences
97 78	66 48	18	66 72	44 46	-2	21 14	32 29	3
74 69	47 45	2	62 22	43 32	9	63 28	43 34	9
58 81	42 49	-7	15 64	29 44	-15	98 42	68 38	30
48 83	40 50	-10	28 37	34 37	-3	86 05	52 18	34
44 43	39 38	1	00 5	10 18	-8	77 94	48 62	-14
73 15	47 29	8	73 7	47 22	25	79 93	49 60	-11
81 73	47 49	-2	56 57	42 42	0	51 29	40 34	6
92 91	60 57	3	04 25	17 33	-16	99 66	70 44	26
79 46	49 39	10	92 53	58 41	17	39 06	37 20	17
63 21	43 32	9	34 94	36 62	-26	17 62	30 43	-13
ΣD		32			-19			87
$\bar{d}$		3.2			-1.9			8.7
ΣD^2		736.00			2229.00			3633.00
$(\Sigma D)^2/10$		102.40			36.10			756.90
$SS = \Sigma D^2-(\Sigma D)^2/10$		633.60			2192.90			2976.10
$s_d^2 = SS/9$		70.40			243.66			330.68
$s_d = \sqrt{s_d^2}$		8.4			15.6			18.2
$s_{\bar{d}} = \sqrt{s_d^2/10}$		2.65			4.83			5.74
$t = (\bar{d}-0)/s_{\bar{d}}$		1.21			-0.39			1.52
$t_{05}s_{\bar{d}} = 2.262s_{\bar{d}}$		5.99			10.93			12.98
C.L. $= \bar{d} \pm t_{05}s_{\bar{d}}$		-2.79 to +9.19			-12.83 to +9.03			-4.28 to +21.68

Table 4.9

Frequency distribution of 250 mean differences, $\bar{d}$,
obtained from samples of 10 differences
obtained as illustrated in Table 4.8 and outlined in section 4.10

Class mark	Frequency
-12	4
-10.5	7
-9	7
-7.5	8
-6	12
-4.5	16
-3	30
-1.5	29
0	33
1.5	21
3	28
4.5	17
6	13
7.5	10
9	8
10.5	4
12	2
13.5	0
15	1
	250

Table 4.9 is a frequency distribution of the resulting 250 mean differences, $\bar{d}$. It will be noticed that the distribution is approximately symmetrical and that 127 of the mean differences are greater than 0 and 123 are less, characteristics associated with a normal distribution. The mean is -0.533, very close to zero.

Tables 4.10 and 4.11 are frequency distributions of the 250 sample variances of 10 differences, s_d^2, and the standard deviations of the 10 differences, s_d, respectively. The forms of these distributions are similar to those given in Tables 4.5 and 4.6 for s^2 and s respectively. However the range of variation is considerably greater than that for individuals. The reason for this is apparent when the possible range of the differences between the variates is considered. The possible range is from (10-70) = -60 pounds to (70-10) = +60 pounds, twice that for the individual observations. The average of the 250 variances, $\overline{s_d^2}$, is 272.7, reasonably close to $2\sigma^2 = 2 \times 144 = 288$. The tables illustrate two important theorems, namely: 1) the variance of random differences, σ_d^2, is twice that of the observations in the original population, and 2) the variance of each sample of differences, s_d^2, is an unbiased estimate of $2\sigma^2$.

In some important cases where the variance of a difference is required, individual differences may not be obtainable. If an s^2 for individuals is available, the appropriate variance, $2\sigma^2$, can be estimated by $2s^2$. From Table 4.5 we have $2s^2 = 2(140.4) = 280.8$, which is very close to the actual variation, observed or theoretical, among the differences. The average of the standard deviation of the differences, $\bar{s}_d$ is 16.04; $\sqrt{\overline{s_d^2}} = \sqrt{272.7} = 16.51$. These compare favorably with $\sigma_d = \sqrt{2\sigma^2} = \sqrt{288} = 16.97$.

It has been stated that $\sigma_{\bar{x}}^2 = \sigma^2/n$ and that $\sigma_d^2 = 2\sigma^2$. From these two theorems, we have that the variance of a difference between means, $\sigma_{\bar{d}}^2$, is equal to $2\sigma^2/n$ when each mean contains n observations. For our data, $\sigma_{\bar{d}} = \sqrt{288/10} = 5.37$; $s_{\bar{d}} = \sqrt{\frac{s_d^2}{n}} = \sqrt{272.7/10} = 5.22$ or $s_{\bar{d}} = \sqrt{\frac{2s^2}{n}} = \sqrt{\frac{280.8}{10}} = 5.30$. Thus, having a value of s^2, we can obtain by formula, estimates of the following important parameters: σ^2, σ_d^2, $\sigma_{\bar{d}}^2$, σ, σ_d and $\sigma_{\bar{d}}$, of which the first three are unbiased. The interrelationships in terms of statistics are shown diagrammatically below:

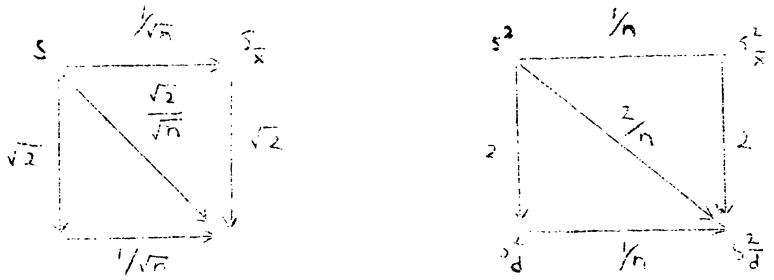
Table 4.10
Frequency distribution of the variances of 10 sample differences (s_d^2)
for 250 random samples of differences based on Table 4.1

Class mark	60	100	140	180	220	260	300	340	380	420	460	500	540	580	620	660	700	740	
Frequency	8	14	24	37	40	34	19	16	12	15	13	7	1	4	2	3	0	1	250

$\overline{s_d^2} = 272.7$ $\quad 2s^2 = 280.8$ $\quad 2\sigma^2 = 288.$

Table 4.11
Frequency distribution of the standard deviations of 10 sample differences (s_d)
for 250 random samples of differences based on Table 4.1

Class mark	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	
Frequency	1	5	4	7	8	17	24	28	29	26	19	13	19	10	16	10	4	4	3	2	1	250



The fact that $s_{\bar{d}}$ can be obtained from s^2 will be seen to be very important when tests of the significance of the difference between two means are discussed.

For each of the 250 samples of 10 differences, t was calculated as $(\bar{d}-0)/(s_{\bar{d}})$. The distribution of these t-values is given in Table 4.12 and is similar to that of Table 4.7 where t was calculated as $t = (\bar{x}-\mu)/(s_{\bar{x}})$. Of these t-values, 118 are positive and 132 are negative; their mean is -0.00013. Fourteen t-values exceeded the 5% as compared to an expected number of 12.5 and 4 exceeded the 1% as compared to an expected number of 2.5.

4.11 Summary of sampling. A summary of the results obtained from the sampling experiment is given in Table 4.13. This summary clearly shows that by sampling it has been possible to demonstrate a number of important characteristics of and theorems concerning normally distributed populations. For example:

i) Means of random samples of n observations are normally distributed with mean, μ, and standard deviation, $\sigma/\sqrt{n}$. (This theorem is approximately true when sampling is from non-normal populations. The degree of approximation depends on the extent of the departure from normality and the sample size.)

ii) Means of differences of random samples of n observations are normally distributed with mean zero and standard deviation $\sqrt{(2\sigma^2/n)}$.

iii) Each random sample provides unbiased estimates of μ, σ^2, $\sigma^2_{\bar{x}}$, σ^2_d, and $\sigma^2_{\bar{d}}$.

lv) The distribution of the statistic $t = (\bar{x}-\mu)/s_{\bar{x}}$ or $t = (\bar{d}-0)/s_{\bar{d}}$ is distributed symmetrically about mean zero. By using a table of t it is possible to make a confidence inference about the population mean from which a sample is drawn or to test a so-called null hypothesis concerning the population mean, e.g. that $\mu = 0$ in the case of differences. Such inferences and tests are based solely on sample data.

Table 4.12

Sample and theoretical values of t, d.f. = 9, $t = \frac{\bar{d}}{s_{\bar{d}}} = 0.$

Interval of t		Sample		Theoretical		
			Percentage	Percentage	Cumulative	
From	To	Frequency	frequency	frequency	One tail	Both tails
------	-3.250	1	0.4	0.5	100.0	
-3.250	-2.821	0	0.0	0.5	99.5	
-2.821	-2.262	7	2.8	1.5	99.0	
-2.262	-1.833	5	2.0	2.5	97.5	
-1.833	-1.383	14	5.6	5.0	95.0	
-1.383	-1.100	13	5.2	5.0	90.0	
-1.100	-0.703	21	8.4	10.0	85.0	
-0.703	0.0	72	28.8	25.0	75.0	
0.0	0.703	61	24.4	25.0	50.0	100.0
0.703	1.100	24	9.6	10.0	25.0	50.0
1.100	1.383	10	4.0	5.0	15.0	30.0
1.383	1.833	7	2.8	5.0	10.0	20.0
1.833	2.262	9	3.6	2.5	5.0	10.0
2.262	2.821	2	0.8	1.5	2.5	5.0
2.821	3.250	1	0.4	0.5	1.0	2.0
3.250	-----	3	1.2	0.5	0.5	1.0
		250	100.0			

Table 4.13

A summary of information from:

i) 500 samples of 10 observations

	$\bar{x}$	$\overline{s^2}$ Divisor: n-1 = 9 n=10		$\overline{s}$ $\sqrt{\overline{s^2}}$		$s_{\bar{x}}$ $s_{\bar{x}}$	$\sqrt{\overline{s^2}/10}$	$\bar{s}/\sqrt{10}$
Sample	39.79	140.42	126.38	11.85	11.48	3.71	3.75	3.63
Population	40.00	144		12		3.79		
	μ	σ^2		σ		$\sigma_{\bar{x}}$		

ii) 250 samples of 10 differences

	$\bar{d}$	$\overline{s_d^2}$ s_d^2	$2s^2$	$\overline{s_d}$ s_d	$\sqrt{s_d^2}$	$\sqrt{2s^2}$	$s_{\bar{d}}$ $s_{\bar{d}}$	$\sqrt{s_d^2/10}$	$\sqrt{2s^2/10}$
Sample	-0.53	272.71	280.8	16.05	16.51	16.74	5.16	5.22	5.30
Population	0	288		16.97			5.37		
	μ	σ_d^2		σ_d			$\sigma_{\bar{d}}$		

iii) t-values

Number of samples	Mean	Number +	Number −	Number beyond without regard to sign $t_{.05} = 2.262$ Obs.	Exp.	Number beyond without regard to sign $t_{.01} = 3.250$ Obs.	Exp.
500	-0.038	248	252	20	25	4	5
250	-0.00013	118	132	14	12.5	4	2.5

CHAPTER V

COMPARISONS INVOLVING TWO SAMPLE MEANS

5.1 Summary. Tests of significance involving the difference between two means and the use of confidence limits to estimate the difference between population means are considered. Procedures are given for cases where the population variance is known, only a sample variance is available, the observations are paired or unpaired, and, in the case of unpaired observations, for equal or unequal sample sizes.

In connection with tests of significance, alternative hypotheses and the choice of a region for rejecting the null hypothesis are discussed as are the two types of error involved. The power of a test is treated.

5.2 Tests of significance. In Chapter 4, section 4.8, for the population of pounds of butterfat from 100 Holstein cows, the means of random samples of 10 observations were compared with the known population mean of 40 pounds by means of the formula $t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$. These t-values were compared with the tabulated t-values for 9 d.f., theoretical values. It was found that approximately 5% of the calculated t-values were equal to or greater than $t_{.05}$ when the signs were ignored. In general, the cumulative percentages for sample and theory are in reasonably good agreement in Table 4.7. This must be so if the theory is correct and reasonably useful. Each of the t-values calculated can be considered as a test of significance and the set of t-values throws light on the totality of conclusions based on such tests.

Consider the making of a single test of significance with t as the statistic. First a rejection region is chosen, i.e. a region such that if the calculated value of t lies in the rejection region, then the hypothesis is to be rejected. For example, a rejection region might be the region of t-values numerically greater than $t_{.05}$. Finally, t is calculated from the observed data. In the calculation of t, a value of μ is substituted dependent upon the hypothesis. If the calculated value of t lies in the rejection region, then we customarily infer that the hypothesis concerning μ was wrong and that the sample mean is too far from the hypothesized population mean to be considered as from the theoretical distribution.

This is not the only inference possible since we could infer that we had an unusual sample from the theoretical distribution. This latter inference would be the correct one for all t-values of Chapter 4 that lay in the rejection region. Shortly, it will be seen that the same type of inference is used when two sample means are compared.

5.3 Null and alternate hypotheses. In experiments in which the significance of the difference between two treatment means is to be evaluated by a test of significance, the hypothesis that there is no difference between the treatments is set up. This is called the null hypothesis. Since the observed means are virtually certain to differ, the hypothesis of no difference applies to the population means estimated by the sample means. The sample means should differ only by random sampling variation if the null hypothesis is true. The null hypothesis can be represented by $\mu_1 = \mu_2$ or by $\mu_1 - \mu_2 = 0$.

If the null hypothesis is not supported by the data, then an alternate hypothesis is inferred. An alternative or a set of alternatives is decided upon when the null hypothesis is set up. For example, a set of alternatives might be that $\mu_1 \neq \mu_2$, i.e. that μ_1 and μ_2 differed, or the set might be that $\mu_1 > \mu_2$, i.e. that μ_1 was greater than μ_2.

The next step is to choose a test statistic, calculate its value for the data obtained, and determine the probability of obtaining a value as discrepant or more discrepant if the null hypothesis is true. If this probability is low, then the null hypothesis is rejected and an alternative accepted, whereas if it is high, the null hypothesis is accepted. Thus, a large sample value of t is regarded as discrepant and the probability of getting a large value when the null hypothesis is true, is small. We reject the null hypothesis. Commonly, it is said that there is a significant difference. When t is the test statistic, it is assumed that there is an underlying normal distribution; more will be said about this when the test is discussed.

The statement regarding the rejection of the null hypothesis is stronger than that regarding the acceptance. The null hypothesis is accepted for lack of evidence to deny it. The experimenter then proceeds on the basis that the difference between population means, if it exists, is small enough to be ignored or that the experiment

was insufficient in size to detect it. Rejection of the null hypothesis is a clear acceptance of the alternate hypothesis which is usually a set of alternatives such as has already been indicated. Regardless of which hypothesis, the null or an alternate, is accepted, the experimenter may be in error and he should consider the consequences of such errors in his choice of a rejection region.

5.4 Levels of significance and the choice of a rejection region. For a test of significance, it is customary to decide upon a level of significance, i.e. a value for the probability of rejecting the null hypothesis when it is true, for reaching a decision to accept or reject the null hypothesis. This is no more than an admission that sampling variation is present and that it is impossible to be correct all of the time when one has only a limited amount of experimental material at his disposal, i.e. when one must make a general inference from a particular example.

In many fields of experimentation, the 5% and 1% levels are customarily used. Thus, if a more discrepant value of the test criterion than that obtained is likely to occur less than 5% of the time but not less than 1% of the time when the null hypothesis is true, the difference is said to be significant and the result marked with a single asterisk in publications. If a more discrepant value of the test criterion than that obtained is likely to occur less often than 1% of the time, the difference is said to be highly significant and is marked with two asterisks in publications. Acceptance of the null hypothesis is often indicated by the letters N.S. or n.s.

The values of 5% and 1% were not acquired by magic or even by mathematics. They seem to have been adequate choices in the field of agriculture where they were first used. In the case of some small-sized experiments, it is possible that the null hypothesis can never be rejected if these levels are required. If an experimenter uses some other level than 5% or 1%, this should be clearly stated.

The choice of a level of significance and a set of alternatives determines the critical region or the region of rejection. Thus, in testing the significance of the difference between two means at the 5% level with t as the test-criterion, one regards large values of t, either positive or negative, as in discord with the null hypothesis.

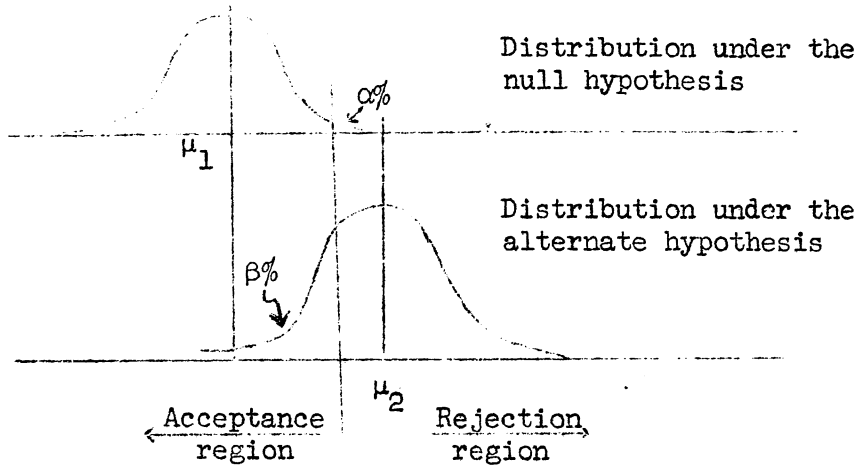
The region of rejection will include all values of t numerically greater than the tabulated value of $t_{.05}$. The probability of obtaining a larger value than $+t_{.05}$ is .025 and of a smaller value than $-t_{.05}$ is .025; the probability of a more discrepant value than $t_{.05}$ is, then, .05, the sum of the probabilities. This paragraph is true when the set of alternatives is simply that the population means are different and there is, consequently, no more reason to look at $\bar{x}_1 - \bar{x}_2$ than there is to look at $\bar{x}_2 - \bar{x}_1$. Such a test is called a two-tailed test.

If the set of alternatives had been $\mu_1 > \mu_2$, then discrepant values of t would be those in which $\bar{x}_1 - \bar{x}_2$ was large and positive. Large negative values would be discrepant but could be attributed only to chance since our set of alternatives does not admit of such values being due to a cause other than chance. (It may be that $\mu_1 < \mu_2$ but that the experimenter simply has no interest in such a result. He simply chooses to ignore it. This might be the case if one were looking for a better variety than some standard rather than simply a variety that differed.) In such a case, the experimenter using a 5% level of significance and t as a test-criterion, would choose his rejection region to contain all positive values of t greater than $t_{.10}$ since the probability of getting a larger positive value of t than $t_{.10}$ is .05. Such a test is referred to as a one-tailed test.

5.5 Two types of error. In making an uncertain inference, there is always the possibility of error. Errors are of two types depending upon whether the null or an alternate hypothesis is true, and it is possible for the experimenter to make either one or the other type of error.

An error of the first kind (Type I) is made when the experimenter rejects the null hypothesis and it is true. The probability of such an error, usually designated by α (Greek alpha), is fixed in advance of the conduct of the experiment, .05 and .01 being common values for α. Thus, if the experimenter is always presented with a sample from the distribution associated with the null hypothesis, he will reject the null hypothesis α% of the time. The upper part of Figure 5.1 shows the situation for a one-tailed test.

Figure 5.1



An error of the second kind (Type II) is made when the experimenter accepts the null hypothesis and the alternate is true. The probability of an error of this type, usually denoted by β, (Greek beta), is determined by the choice of α and the separation between μ_1 and μ_2. This can be seen in the lower part of Figure 5.1. Thus, if the experimenter is always presented with a sample from the distribution associated with the alternate hypothesis, he rejects this hypothesis when he accepts the null hypothesis; this amounts to $\beta\%$ of the time. The concept of type II error is especially important in determining the sample size necessary to detect a difference of a stated magnitude. This, in turn, should tell the experimenter whether or not he has sufficient resources to conduct an adequate experiment.

When both kinds of error are considered, it is apparent that a reduction in the probability of a type I error must be accompanied by an increase in the probability of a type II error. This is readily seen from Figure 5.1, since a decrease in α requires the moving of the line separating the acceptance and rejection regions to the right with a consequent increase in the size of β. This points to the need of considering the seriousness of the different types of error in choosing a level of significance. If it is a serious error to fail to detect a real difference, i.e. if accepting the null hypothesis when it is false is serious, then β should be small. Since the population means are fixed, this can be done only by moving the

line separating acceptance and rejection regions, to the left. This results in an increase in α, i.e. the probability of rejecting the null hypothesis when it is true must be increased. In an ideal situation, both α and β are fixed in advance. This determines the required sample size.

Closely related to type II error is the power of the test. The power of a test is its ability to detect the alternate hypothesis when it is true. This is seen to be associated with the area under the lower curve of Figure 5.1 for the rejection region. If the alternate hypothesis is rejected with probability β when it is true, then it is accepted with probability $1-\beta$ when it is true. It is obviously desirable to have this probability high.

5.6 Basis of the test of two means. In Chapter 4, section 4.7, we learned that the variance of sample means was given by σ^2/n where σ^2 was the variance of the parent population. Thus, two sources are available for estimating σ^2, viz. means and individuals. Basically, the test of significance for the difference between two sample means consists in determining the ratio of the two estimates of the population variance. In particular, we have

$$\frac{\text{Estimate of } \sigma^2 \text{ from means}}{\text{Estimate of } \sigma^2 \text{ from individuals}}$$

or its square root as the test criterion. This ratio is usually considered to be distributed as t^2, in which case it can be compared to tabulated values of t^2 for a number of percentage points, or its square root can be compared with tabular values of t. In either case, degrees of freedom are involved.

If the ratio is not distributed as t, it will be compared with an approximation given by Cochran and Cox.

The form of the test criterion that is generally used is $\frac{\bar{d}}{s_{\bar{d}}}$ where $\bar{d} = \bar{x}_1 - \bar{x}_2$, the difference between the two sample means, and $s_{\bar{d}}$ is the standard deviation appropriate to a difference between two means selected at random from populations with a common mean. That $\bar{d}$ is essentially a standard deviation was shown in Chapter 2, section 2.10. Calculation of $s_{\bar{d}}$ depends on i) whether or not the two populations have a common variance, ii) whether the value of the σ^2's (or the common σ^2) are known or estimated, iii) whether or not the samples from the two populations are of the same size, and iv) whether or not the variates are paired. The choice of a rejection

region depends upon i) the level of significance chosen and ii) whether a one- or two-tailed test is involved. The comparison of a sample mean with a population mean has already been discussed and it will be seen that a test involving two means is directly related.

The following discussion will be concerned with two-tailed tests. The modification necessary for a one-tailed test will be indicated toward the end of the chapter.

5.7 Comparison of sample means from possibly different populations for unpaired variates. Given a sample from each of two populations, we desire to ascertain whether the population means are the same. Let μ_1 and μ_2 denote the two population means, σ_1^2 and σ_2^2 the population variance, $\bar{x}_1$ and $\bar{x}_2$ the sample means, s_1^2 and s_2^2 the sample variances estimating σ_1^2 and σ_2^2 respectively, and n_1 and n_2 the number of variates in each sample.

Case 1. To test the hypothesis that the two population means are equal, i.e. that $\mu_1 = \mu_2$, given a sample from each population. Assume that the population variances are the same, i.e. that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, say. In the test criterion, $s_{\bar{d}}$ is an estimate of $\sigma_{\bar{d}}$ and subject to sampling variation. Consequently, the distribution of the criterion is different for each value of the number of degrees of freedom on which s is based. The estimate used should make the most efficient use of the data and is customarily calculated by pooling the sums of squares from each sample and dividing by the pooled d.f. The criterion is distributed as t when the underlying distributions are normal but considerable departures from normality may not seriously affect it.

This problem was first studied by Student in 1908. He prepared the distribution of a related statistic which he called Z. Later, R. A. Fisher worked out the exact distribution of t on which the tables are based.

a) The test for $n_1 \neq n_2$. The test criterion is

$$t = \frac{(\bar{x}_1 - \mu_1) - (\bar{x}_2 - \mu_2)}{s_{\bar{d}}}$$

which becomes

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{d}}}$$

for the chosen null hypothesis. In the more general test, the difference $\mu_2 - \mu_1$ may be set equal to any desired value and the

test is similar to that of Chapter 4 where a sample mean was tested against a hypothesized population mean.

$$s_{\bar{d}} = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{s^2\left(\frac{n_1 + n_2}{n_1 n_2}\right)}$$

where s^2 is the pooled mean square of the two samples.

The numerical procedure is given in the following example:

C. J. Watson et al. - Digestibility studies with Ruminants XII - The Comparative Digestive Powers of Sheep and Steers

Sci. Agr. 28: 357-374, 1948

Coefficients of digestibility in percent of dry matter - feed corn silage

	X_1	X_2
	Sheep	Steers
	57.8	64.2
	56.2	58.7
	61.9	63.1
	54.4	62.5
	53.6	59.8
	56.4	59.2
	53.2	
ΣX:	393.5	367.5
ΣX^2:	22174.41	22535.87
$\bar{x}$	56.21	61.25

$$\Sigma x^2_1 = \Sigma X_1^2 - (\Sigma X_1)^2/n_1 = 22174.41 - 22120.32 = 54.09$$

$$\Sigma x^2_2 = \Sigma X_2 - (\Sigma X_2)^2/n_2 = 22535.87 - 22509.37 = 26.50$$

$$s^2 = \frac{\Sigma x_1^2 + \Sigma x_2^2}{(n_1-1)+(n_2-1)} = \frac{54.09 + 26.50}{6 + 5} = 7.32,$$ the pooled estimate of a common σ^2.

$$s_{\bar{d}} = \sqrt{s^2\frac{(n_1 + n_2)}{n_1 n_2}} = \sqrt{7.32\frac{(7 + 6)}{42}} = \sqrt{2.27} = 1.51,$$

the standard deviation appropriate to the difference between the sample means.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{d}}} = \frac{56.21 - 61.25}{1.51} = \frac{5.04}{1.51} = 3.33^{**},$$ d.f. = $(n_1-1) + (n_2-1) = 11$.

For the 95% confidence interval, calculate $\bar{d} \pm t_{.05}\, s_{\bar{d}}$

$= 5.04 \pm 2.201(1.51) = 5.04 \pm 3.32$ $\therefore$ $l_1 = 1.72$; $l_2 = 8.36$

The 95% confidence interval for the difference $\mu_2-\mu_1$ is also shown.

b) When $n_1 = n_2 = n$, say, the same procedure is applicable but the arithmetic can be simplified for $s^2(\frac{n_1+n_2}{n_1 n_2})$ reduced to $\frac{2s^2}{n}$. The degrees of freedom are 2(n-1).

The following example demonstrates the procedure:

R. H. Ross and C. B. Knodt - The effect of supplemental vitamin A upon growth, blood plasma, carotene, vitamin A, inorganic calcium, and phosphorus content of Holstein heifers.

Jour. Dairy Sci. 31:1062-1067, 1948

Gain in weight

X_1 Control	X_2 Vit. A.
175	142
132	311
218	337
151	262
200	302
219	195
234	253
149	199
187	236
123	216
248	211
206	176
179	249
206	214
ΣX : 2627	3303
ΣX^2: 511807	817583
$\bar{x}$: 187.6	235.9

$n_1 = n_2 = n = 14$

$\Sigma x_1^2 = \Sigma X_1^2 - (\Sigma X_1)^2/n = 511807 - 492938 = 18869$

$\Sigma x_2 = \Sigma X_2^2 - (\Sigma X_2)^2/n = 817583 - 779272 = 38311$

$s^2 = \frac{\Sigma x_1^2 + \Sigma x_2^2}{2(n-1)} = \frac{57180}{26} = 2199.$

$s_{\bar{d}} = \sqrt{\frac{2s^2}{n}} = \sqrt{\frac{2(2199)}{14}} = 17.7$

$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{d}}} = \frac{187.6 - 235.9}{17.7}$

$= \frac{48.3}{17.7} = 2.73*$

d.f. = 2(n-1) = 26; $(t_{.01} = 2.78)$.

For the 95% confidence interval, calculate

$\bar{d} \pm t_{.05} s_{\bar{d}} = 48.3 \pm 2.056(17.7) = 48.3 \pm 36.4$

Hence, $l_1 = 11.9$ and $l_2 = 84.7$

For the 99% confidence interval, calculate

$\bar{d} \pm t_{.05} s_{\bar{d}} = 48.3 \pm 2.779(17.7) = 48.3 \pm 49.2$

Hence, $l_1 = -0.9$ and $l_2 = 97.5$

The 95% and 99% confidence intervals are given for the difference $\mu_2 - \mu_1$.

When σ^2 is known, the criterion becomes

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma^2 \frac{n_1 + n_2}{n_1 n_2}}}$$

which is compared with the tabulated values in the last line of the t-table. These values are taken directly from tables of the normal distribution. It is seen, then, that a normal approximation is

used when dealing with large samples.

5.8 The linear model for the case of section 5.7. The linear model for the preceding case states that any observation is one made to obtain information on the appropriate population mean. When several observations are made, it is possible to obtain an estimate of the population variance. Denote by X_{ij} the j-th observation on the i-th population, with i equal to 1 or 2 and $j = 1, 2, \ldots, n_1$ when $i = 1$ and $j = 1, 2, \ldots, n_2$ when $i = 2$. Then the model states that

$$X_{ij} = \mu + \tau_i + \epsilon_{ij} \quad (\tau \text{ is Greek tau.})$$

In terms of the notation used in the previous discussion, $\mu_1 = \mu + \tau_1$ and $\mu_2 = \mu + \tau_2$. The present notation simply states that any observation is composed of a general mean, μ, to which has been added a contribution, τ_i, due to the specific population or treatment, and an error, ϵ_{ij}, associated with the observation itself. For arithmetic convenience, we set $\Sigma\tau_i = 0$ or $\tau_1 = -\tau_2$ in this case.

In the process of obtaining s^2, we calculate $\Sigma(X_{1j}-\bar{x}_1)^2$ which is a sum of squares associated with ϵ_{1j}'s only since $\mu + \tau_1$ is common to all these observations. Similarly we calculate a sum of squares associated with ϵ_{2j}'s only. These are pooled on the assumption that the ϵ's are from a single population. We note that no contribution due to the τ's enters our estimate of σ^2.

In the process of calculating $\bar{x}_1 - \bar{x}_2$, a difference between $\mu + \tau_1$ + an average of ϵ's and $\mu + \tau_2$ + an average of ϵ's is involved. As has been shown, this difference is equivalent to a standard deviation and has a contribution due to the τ's as well as the ϵ's. Thus, if we obtain a large value of t and do not wish to attribute this to a chance happening, we conclude that it is due to the contribution of the τ's and that the evidence does not support the null hypothesis.

5.9 Comparison of sample means from possibly different populations for paired variates. Frequently there arises the situation where the variates are paired. For example, two rations may be com red using two animals from each of 10 litters of swine, one animal from each litter being assigned at random to a ration, and the other animal assigned to the other. Or, the percentage of oil in two soybean varieties grown in pairs of plots at 12 locations may be compared. In this instance, the two observations, one for each variety for a particular station, constitute a pair. If the members

of the pair tend to be positively correlated, that is the members of a pair tend to be more alike than members of different pairs, an increase in the precision of our test is possible as compared to a non-paired experiment. This information is utilized in the test in that the part of the variation among individuals treated alike which is common to both members of a pair but varies from pair to pair, can be removed from the experimental error. Actually the experimental error is based upon the variation of the pair differences. The degrees of freedom for estimating σ^2 are one less than the number of pairs. If pairing had not been used, the number of degrees of freedom available for estimating σ^2 would have been twice as large. Thus if the precision of our experiment is to be increased, the reduction in the variance due to pairing must more than compensate for the loss in precision due to fewer degrees of freedom being available for estimating σ^2.

As before, we calculate $t = \frac{\bar{d}}{s_{\bar{d}}}$ where $s_{\bar{d}} = \sqrt{\frac{\Sigma(X_1-X_2)^2-\frac{[\Sigma(X_1-X_2)]^2}{n}}{n(n-1)}}$ and is based upon n-1 d.f. since there are n pairs of observations.

The mean square among the differences in the above formula, viz. $s_d^2 = \frac{\Sigma(X_1 - X_2)^2 - \frac{[\Sigma(X_1 - X_2)]^2}{n}}{n-1}$, is clearly analogous to the working formula for sums of squares, viz. $\frac{\Sigma X^2 - (\Sigma X)^2/n}{n-1}$.

The mean square for the differences, when written in definition form analogous to $\frac{\Sigma(X - \bar{x})^2}{n-1}$, is $\frac{\Sigma [(X_1 - X_2) - (\bar{x_1} - \bar{x_2})]^2}{n-1}$ $= \frac{\Sigma [(X_1 - \bar{x}_1) - (X_2 - \bar{x}_2)]^2}{n-1}$. Let $X_1 - \bar{x}_1 = x_1$ and $X_2 - \bar{x}_2 = x_2$ and we have $\frac{\Sigma(x_1 - x_2)^2}{n-1} = \frac{\Sigma x_1^2}{n-1} + \frac{\Sigma x_2^2}{n-1} - \frac{2\Sigma x_1 x_2}{n-1} = s_1^2 + s_2^2 - \frac{2\Sigma(x_1 x_2)}{n-1}$

Thus s_d^2 is equal to the sum of the mean squares of X_1 and X_2 less twice $\Sigma(x_1 x_2)$, called the covariance, a quantity which we shall hear more about later. It is sufficient to say here that if high values of X_1 are associated with high values of X_2, then the covariance will be a positive quantity. Thus, the more similar the members of a pair are as compared to members of different pairs, the larger will be the covariance term and the greater the reduction

in $s^2_{\bar{d}}$ as compared to the case of no pairing.

The case of paired variates has, then, been reduced to the case of testing the hypothesis that the mean of the differences is a specified number, often taken to be zero. An example is worked below.

R. W. Shuel Some factors affecting nectar secretion in red clover. Plant Physiol. 27:95-110, 1952

Sugar concentration of nectar in 1/2 heads of red clover kept at different vapor temperatures for 8 hours.

Vapor pressure 4.4 Mm. Hg. X_1	9.9 Mm. Hg. X_2	$X_1 - X_2$
62.5	51.7	10.8
65.2	54.2	11.0
67.6	53.3	14.3
69.9	57.0	12.9
69.4	56.4	13.0
70.1	61.5	8.6
67.8	57.2	10.6
67.0	56.2	10.8
68.5	58.4	10.1
62.4	55.8	6.6
ΣX: 670.4	561.7	108.7
		1226.07
$\bar{x}$: 67.0	56.2	10.8

$$s^2_{\bar{d}} = \frac{\Sigma(X_1-X_2)^2 - \Sigma(X_1-X_2)^2/n}{n(n-1)} = \frac{1226.07 - 1181.57}{(9)(10)} = 0.4944$$

$$s_{\bar{d}} = .703$$

$$t = \frac{\bar{d}}{s_{\bar{d}}} = \frac{10.8}{.703} = 15.4^{**}, \text{ for 9 d.f.}$$

(Note that $\Sigma(X_1-X_2) = \Sigma X_1 - \Sigma X_2$ and that $\overline{x_1-x_2} = \bar{x}_1 - \bar{x}_2$.)

The 99% confidence interval for the population mean difference is calculated as $\bar{d} \pm t_{.01}s_{\bar{d}} = 10.8 \pm 3.3(.703)$.

Hence $l_1 = 8.5$ and $l_2 = 13.1$

The hypothesis tested was that the mean of the population of differences was zero. The test criterion is distributed as t when the assumption that differences are normally distributed is correct and the null hypothesis true. Tabulated $t_{.01}$ for 9 d.f. is 3.3.

When σ^2 is known, calculated t is compared with tabulated t for the last line of the t-table. These values are the same as those to be found in tables of the normal distribution. For large sample sizes, the normal approximation is seen to be very little different from the more exact value of t based on the number of d.f.

5.10 The linear model for the paired comparison. The linear model is stated in the equation

$$X_{ij} = \mu + \tau_i + \rho_j + \epsilon_{ij}$$

where X_{ij} is the observation on the i-th treatment for the j-th pair for $i = 1, 2$ and $j = 1, 2, \ldots, n$. It is seen that the model admits of a different population mean for each observation but that these means are closely related due to their manner of construction. It is because of this relation that we are able to estimate σ^2, the variance of the ϵ's. To see this clearly, set up a table as follows:

X_{1j}	X_{2j}	$X_{1j} - X_{2j}$
$\mu + \tau_1 + \rho_1 + \epsilon_{11}$	$\mu + \tau_2 + \rho_1 + \epsilon_{21}$	$(\tau_1 - \tau_2) + (\epsilon_{11} - \epsilon_{21})$
.....	.....	.....
$\mu + \tau_1 + \rho_n + \epsilon_{1n}$	$\mu + \tau_2 + \rho_n + \epsilon_{2n}$	$(\tau_1 - \tau_2) + (\epsilon_{1n} - \epsilon_{2n})$

It is immediately seen that the differences of the last column have a variation associated only with differences (algebraic sums) of ϵ's since the difference $(\tau_1 - \tau_2)$ is a constant in all differences. The variance of the sample differences is an estimate of $2\sigma^2$. The numerator of the test criterion involves the sum of the differences and, consequently, has a contribution from the difference between treatments, if such a difference is present, in addition to a contribution due to error. If, then, the numerator is much larger than the denominator the largeness is customarily attributed to a real treatment difference rather than to an unusual chance event.

This test has one important property not possessed by the previous tests involving the hypothesis of a difference between two treatment means. Theory tells us that the algebraic sum of normally distributed variables is normally distributed. The application of this to the present case is to the effect that the differences are normally distributed provided the errors are, regardless of whether or not the ϵ_1's and the ϵ_2's have a common variance.

A value of pairing not previously mentioned concerns the scope of inference. It is seen that the variation from pair to pair can be large. If we deliberately make this variation large, we widen the scope of our inference. Thus our pairs of swine come from many litters involving different sires and, possibly, different breeds; our soybean varieties were grown at different locations rather than a single location; our inference is broadened.

5.11 Unpaired observations and unequal variances. Given a sample from each of two populations where $\sigma_1^2 \neq \sigma_2^2$, i.e. with unequal variances. It is desired to test the hypothesis that $\mu_1 = \mu_2$ using the sample estimates of the variances.

Here, the ratio $\bar{d}/s_{\bar{d}}$ is no longer distributed as t. However, a sufficiently accurate approximation given by Cochran and Cox for determining significant values of t for a given significance level is available. The test criterion is $\bar{d}/s_{\bar{d}}$ where

$$s_{\bar{d}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Note that the standard deviation of a difference does not involve the pooling of the sums of squares and d.f. as was the case when s_1^2 and s_2^2 were estimates of a common variance.

The value of t to be judged significant is calculated as

$$t' = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2} \quad \text{where } w_1 = \frac{s_1^2}{n_1}, \; w_2 = \frac{s_2^2}{n_2},$$

and t_1 and t_2 are the values of tabulated t for $n_1 - 1$ and $n_2 - 2$ respectively for the chosen level of significance. The sample value of the criterion is compared with t'. This approximation errs slightly on the conservative side in that the value of t' required for significance may be slightly too large.

It is seen that t' must always lie between the tabulated t values for $n_1 - 1$ and $n_2 - 1$ d.f. Hence the calculation is needed only for those cases where the difference is on the borderline. When $n_1 = n_2 = n$, say, t' is the tabular t for $n - 1$ d.f.

The following exemplifies the procedure.

W. Roules. Physical properties of mineral soils of Quebec.
Canada Journ. Res. 16: 277-287, 1938

% fine gravel in surface soils

good soil	poor soil
5.9	7.6
3.8	0.4
6.5	1.1
18.3	3.2
18.2	6.5
15.1	4.1
7.6	4.7
ΣX: 76.4	27.6
ΣX^2: 1074.60	150.52
$\bar{x}$: 10.91	3.94

$$\Sigma x_1^2 = \Sigma X_1^2 - (\Sigma X_1)^2/n_1 = 1074.60 - 833.85 = 240.75$$

$$s_1^2 = \frac{\Sigma x_1^2}{n_1 - 1} = \frac{240.75}{6} = 40.12$$

$$\Sigma x_2^2 = \Sigma X_2^2 - (\Sigma X_2)^2/n_2 = 150.52 - 108.82 = 41.70$$

$$s_2^2 = \frac{\Sigma x_2^2}{n_2 - 1} = \frac{41.70}{6} = 6.95$$

$$s_{\bar{d}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_1}} = \sqrt{\frac{40.12}{7} + \frac{6.95}{7}} = \sqrt{6.72} = 2.59$$

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{d}}} = \frac{10.91 - 3.94}{2.59} = \frac{6.97}{2.59} = 2.69$$

Compare t' with tabular t for 6 d.f. = 2.45 at 5%

The 95% confidence interval is calculated from

$$\bar{d} \pm t_{05} s_{\bar{d}} = 6.97 \pm 2.45(2.59) = 6.97 \pm 6.35$$

Hence $l_1 = 0.62$; $l_2 = 13.32$

When the population variances are known, the same criterion is used but the sample value of the criterion is compared with the values given in the last line of the t-table or directly with values of normal deviates. In this case, the test is not an approximation if the underlying distributions are normal.

5.12 Testing the hypothesis of equality of variances. In section 5.4, the choice of a critical or rejection region was seen to depend upon the set of alternative hypotheses. The use of one- and two-tailed tests was discussed. The tests just discussed have been treated from the point of view of the two-tailed test since the set of alternate hypotheses was implied to be simply that a difference existed. For one-tailed test of the hypotheses discussed, the test criterion is the same but therejection region is chosen as stated in section 5.4.

It has already been pointed out that the basis of the test criterion is the comparison of two variances. This was, perhaps, more obvious from the square of the criterion than from the criterion itself. This test implies the possibility of generalization so that any two variances, regardless of the number of d.f. in either, may be compared under the hypothesis that they are sample variances from populations with a common variance.

Such a test would be available for the purpose of deciding whether or not it was legitimate to pool variances as was done in testing the hypothesis of the equality of population means when samples were drawn without pairing from two populations. An appropriate criterion for testing this hypothesis is called F. Tabulated values of F are required for every possible pair of d.f., i.e. the test criterion, F, has a distribution for each possible pair of d.f. Values for quite a few pairs of d.f. and for .05 and .01 levels of significance are given in table 5.1.

Consider testing the hypothesis that $\sigma_1^2 = \sigma_2^2$ against the set of alternatives $\sigma_1^2 \neq \sigma_2^2$. Determine s_1^2 and s_2^2 and obtain

$$F = \frac{\text{the larger } s^2}{\text{the smaller } s^2} .$$

This value of F is then compared with tabular values of F given in Table 5.1, where the degrees of freedom for the larger mean square are given across the top of the table and those for the smaller along the side. For the set of alternatives $\sigma_1^2 = \sigma_2^2$, the tabulated 5% and 1% values are 10% and 2% levels respectively. If the calculated F is larger than $F_{.05}$, we claim significance at the 10% level, and if larger than $F_{.01}$, we claim significance ar the 2% level. Values for other levels are available elsewhere.

This test is a two-tailed test since we do not specify which σ^2 is expected to be larger.

An example of this test is given below for the example of section 5.11. Calculate F = 40.12/6.95 = 5.77 for 6 d.f. in both numerator and denominator. Tabulated F values are 4.28 and 8.47 so our test shows significance at the 10% level but not at the 2% level. It would appear safe to conclude that the variances did not estimate a common population value.

The F-tables are tabulated for convenience in making one-tailed tests since the associated alternatives are more common. Thus, in the t-tests of this chapter, it was seen that the numerator was expected to be larger when the null hypothesis was false, i.e. the denominator variance had to be large to deny the null hypothesis. If you square any tabulated value of t, you will find the square in the columns of the F-table headed by 1 d.f. and opposite the appropriate number of d.f. A test using two-tailed t is seen to be equivalent to one using one-tailed F.

5.13 Confidence limits involving the difference between two means. If is often of interest to establish a range within which we can state that the true difference $\mu_1 - \mu_2$ lies. Since it is not possible to do this with certainty, a measure of the reliability of our conclusion must be given. This is the same problem of estimation that we discussed in chapter 2 in estimating a range about $\bar{x}$ within which we stated that μ lay. We were able to say what proportion of our statements were correct but not able to say which were the correct ones unless we knew μ.

Denoting $\mu_1 - \mu_2$ by δ, we solve for δ in each of the equations $\pm t = \frac{\bar{d} - \delta}{s_{\bar{d}}}$ where the tabulated value of t associated with the desired probability level is substituted for t. For example, for a level of .05, we have

$$l_1 = \bar{d} - t_{.05}s_{\bar{d}} \quad \text{and} \quad l_2 = \bar{d} + t_{.05}s_{\bar{d}} \ .$$

If a test of significance is applied and t is less than $t_{.05}$, then the confidence interval will include zero; if t is greater than $t_{.05}$, then the confidence interval will not include zero. Examples are shown with the previous problems.

5.14 Sample size and the detection of differences. Many experiments are conducted where a control and a treatment constitute two treatments such that the experimenter is interested in detecting a difference of a stated size only if the treatment is superior to the control. A one-tailed t-test is appropriate.

Suppose it is desired to compare a new sugar-beet variety with a standard. It is desired to detect a real difference of 50 pounds, (i.e. a difference of 50 pounds between the true means for the two populations in favor of the new variety), if it exists, for a stated plot size with probability .80. The test of the significance of a difference is to be made with the significance level set at 5%, all in one tail.

Consider figure 5.2. The two distributions A and B represent the populations associated with the mean sample differences between the control and the new variety when there is no real difference, and when the real difference is 50 pounds in favor of the new variety, respectively. The exact location of means for the sugar-beet populations is unimportant; only the separation is important. Now we need to determine the distance d such that 5% of the area of curve A lies to the right of d and 80% of the area under curve B lies to the right of d. Then using the point d as the dividing line for making the decision between A and B, the experimenter will declare in favor of B 5% of the time when A is the correct population and will declare in favor of B 80% of the time when it is the correct population. The problem is to choose the sample size n, i.e. the number to be observed for each treatment, such that the two distributions of sample mean differences will have variances permitting a decision between A and B to be made by use of d, in such a way as to have the specified areas on either side of d. This can be done by obtaining an expression for d from knowledge of each population. From A and the test of significance between two treatment means distributed about a population mean of zero, we obtain

$$d = 1.65\sqrt{\frac{2}{n}}\,\sigma.$$

From A and the requirement that we detect a difference of 50 pounds, if it exists, in 80% of our experiments, we have

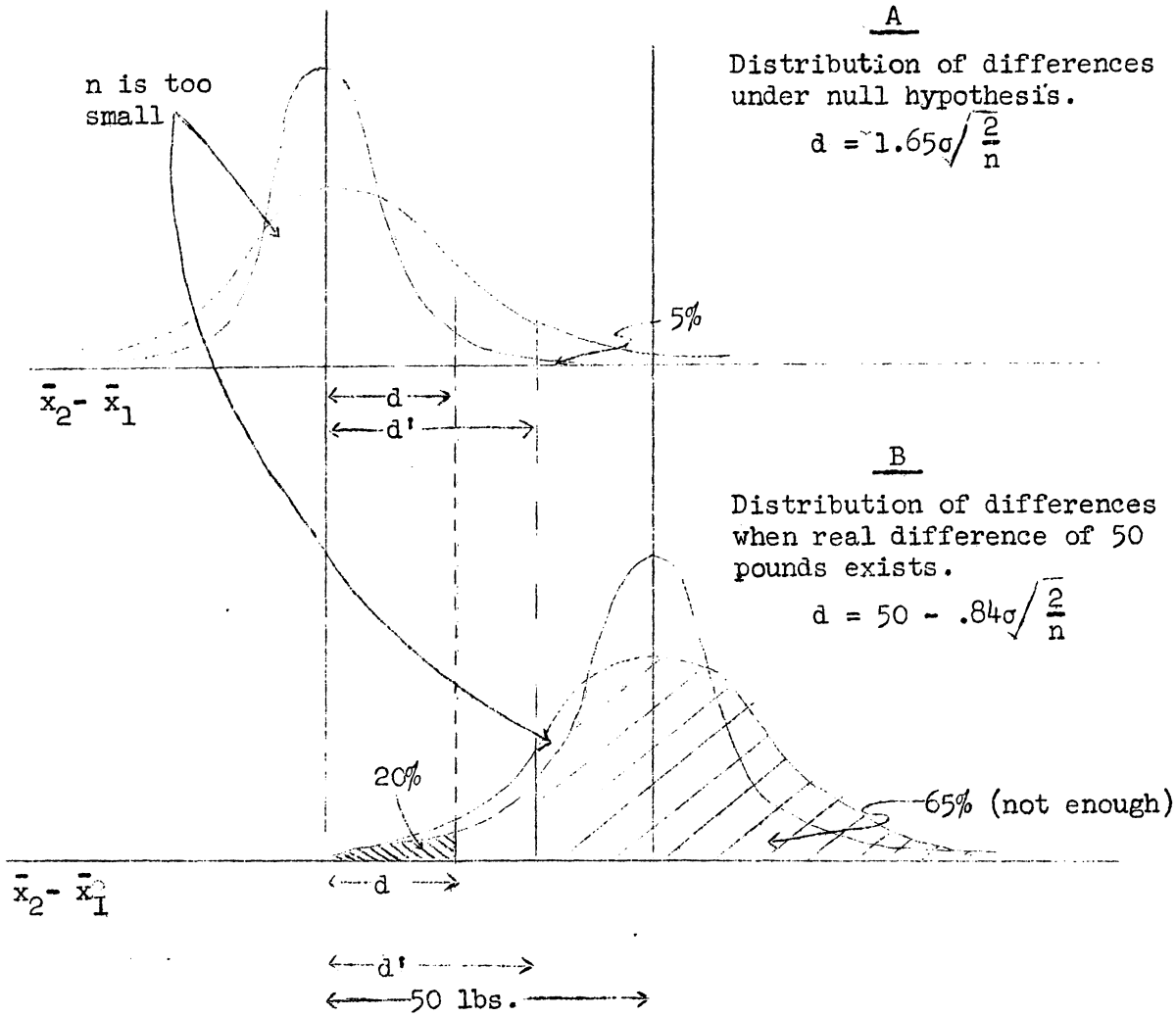
$$d = 50 \text{ (pounds)} - .84\sqrt{\frac{2}{n}}\,\sigma$$

since 20% of the area under a normal curve with zero mean lies to the left of -.84. Thus

$$d = 1.65\sqrt{\frac{2}{n}}\,\sigma = 50 - .84\sqrt{\frac{2}{n}}\,\sigma,$$

$$\sqrt{\frac{2}{n}}\,\sigma = \frac{50}{1.65 + .84} =$$

Figure 5.2



(σ^2 = variance of individuals)

The n that determined d' was too small for it detected a real difference of 50 pounds only 65% of the time. A larger n decreased $\sigma/\sqrt{\frac{2}{n}}$, moved d closer, and increased area under the lower curve to the right of d.

and $$n = \frac{2\sigma^2 (2.49)^2}{50^2} = .00496\sigma^2.$$

A value of σ^2 is required to complete the calculations. If only an estimate is available, this complicates the problem to some extent and leads to higher values of n. The value of n permits differences larger than 50 pounds to be detected with a probability greater than .80 whereas a small difference will be detected less often.

For more complete information on the subject of sample size, see Cochran and Cox, Experimental Designs, table 2.1, pages 20 and 21, and figure 7.1 after page 260, Paulson and Wallis, chapter 7 of Techniques of Statistical Analysis. The latter is a monograph for determining sample size when two percentages are involved.

CHAPTER 8

ANALYSIS OF VARIANCE I

8.1 Summary. In this chapter, the procedure for testing the hypothesis of the equality of k population means is given. Presentation of the results in the form of an analysis of variance table is treated showing the additivity of sums of squares and d.f. The use of a least significant difference, confidence limits, a linear model, and variance components are discussed. The underlying assumptions and difficulties arising when these assumptions are false are considered. Sampling and experimental errors are dealt with.

8.2 Data with a single classification. The analysis of variance for two groups. In section 7 of Chapter 5, data on coefficients of digestibility were given. A single classification was used with the data, namely that of type of animal, for which there were two categories. It is clear, then, that the observations within each category are assumed to be a random sample from a single population. The populations associated with each category are assumed to have a common variance and the hypothesis of a common mean is to be tested. In Chapter 5, Student's t-test was used to test the hypothesis of a common mean for the data now being discussed.

In section 6 of Chapter 5, it was stated that the basis of a test of the equality of two population means using sample data was a comparison of estimates of the variance σ^2, one available from means and the other from individuals. The variance of two means, when based on equal numbers of observations, is given by

$$\bar{x}_1^2 + \bar{x}_2^2 - \frac{(\bar{x}_1 + \bar{x}_2)^2}{2}$$

and is an estimate of σ^2/n. To estimate σ^2, it must be multiplied by n. This gives

$$n\bar{x}_1^2 + n\bar{x}_2^2 - \frac{n(\bar{x}_1 + \bar{x}_2)^2}{2}$$

$$= \frac{n^2\bar{x}_1^2}{n} + \frac{n^2\bar{x}_2^2}{n} - \frac{n^2(\bar{x}_1 + \bar{x}_2)^2}{2n}$$

$$= \frac{(\Sigma X_1)^2}{n} + \frac{(\Sigma X_2)^2}{n} - \frac{(\Sigma X_1 + \Sigma X_2)^2}{2n}$$

This quantity is usually called the sum of squares attributable to means or the sum of squares for means. Since there is a single d.f., it is also the variance. It is also easy to see that the last line of the algebra, if the n is ignored, gives the sum of squares and the variance of the two sums. Since this would be an estimate of $n\sigma^2$, the divisor n would be required if it were desired to estimate σ^2, which is the case. Also the last line of the preceding algebra implies the usual computational form which is especially convenient for machine calculations. In practice, calculate the sum of the squared sums and divide the result by n rather than divide each squared sum by n. Note that the divisors of the squares or the sums of squares are the numbers of observations entering the sums to be squared.

The sum of squares used to calculate the variance of individuals is

$$\Sigma x_1^2 + \Sigma x_2^2 = \Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n} + \Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n} .$$

Addition of the sums of squares for means and for individuals gives

$$\frac{(\Sigma X_1)^2}{n} + \frac{(\Sigma X_2)^2}{n} - \frac{(\Sigma X_1 + \Sigma X_2)^2}{2n} + \Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n} + \Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n}$$

$$= \Sigma X_1^2 + \Sigma X_2^2 - \frac{(\Sigma X_1 + \Sigma X_2)^2}{2n}$$

which is immediately recognizable as the sum of squares of all the observations.

This property of <u>additivity of sums of squares</u> is characteristic of well-planned and executed experiments. It leads to certain shortcuts in arithmetic. For example, the usual calculation procedure for the analysis of variance for such data consists of finding the sum and the sum of squares for all observations in a single operation on a calculating machine. The sum is squared and divided by the total number of observations to give the <u>correction factor</u>. From these quantities, the <u>total sum of squares</u> is obtained. In turn, the sum of squares attributable to means, also called the <u>between</u> or <u>among groups sum of squares</u> or the <u>treatment sum of squares</u>, is calculated from the group subtotals. These are usually wanted by the experimenter since he will also wish to observe the means when drawing con-

clusions from his experiment. The correction factor here is the same as for the total sum of squares so has already been calculated and the sum of the subtotals, obtained in the computing machine along with their sum of squares, is the total. This gives a check on the arithmetic. The sum of squares used to calculate the variance of individuals, also called the within groups sum of squares, residual sum of squares, or error sum of squares, is generally obtained by subtraction of the between groups sum of squares from the total sum of squares. We have seen how this can be done directly for the one-way classification but there is no convenient direct method for most higher-order classifications.

The numerical results are usually presented in an Analysis of Variance Table such as Table 8.1. Note that the d.f. as well as the sums of squares are additive. It is customary to calculate the mean squares where checked, since these are the estimates of σ^2 to be compared. A third estimate of σ^2 is available from the total line when the null hypothesis is true; however, the two in the body of the table are independent of each other. Independence is a requirement for the F-test to be valid.

Table 8.1

Analysis of Variance (One-way classification)

Source of Variation	d.f.	Sum of Squares	Mean Square
Treatments	1	$\frac{(\Sigma X_1)^2 + (\Sigma X_2)^2}{n} - \frac{(\Sigma X_1 + \Sigma X_2)^2}{2n}$	✓
Residual	2(n-1)	By subtraction	✓
Total	2n - 1	$\Sigma X_1^2 + \Sigma X_2^2 - \frac{(\Sigma X_1 + \Sigma X_2)^2}{2n}$	

For the gain-in-weight data of Chapter 5, section 7, the analysis of variance is

Table 8.2

Source of Variation	d.f.	Sum of Squares	Mean Square	F
Treatments	1	16,320.5	16,320.5	7.42*
Residual	26	57,180.2	2,199.2	
Total	27	73,500.7		

The tabulated F-values for 1 and 26 d.f. are 4.22 and 7.72 at the .05 and .01 probability levels respectively. The F-tables, for the analysis of variance, are always entered with the numerator d.f. along the top of the table and the denominator d.f. along the side. This is because the set of alternative hypotheses admits only that treatment differences exist and consequently increase the estimate of variance so that the test is performed with the treatment sum of squares in the numerator. If treatment sum of squares is less than the residual, then the result is declared non-significant no matter how small the ratio.

In order that the F-test be valid, it is necessary to make several assumptions, namely that the underlying distributions be normal and that they have a common variance. It is known that the F-test can stand considerable departure from normality before it is seriously affected at the customary probability levels. These assumptions are not required for the arithmetic to be valid. Note that the same conclusion is drawn as with the t-test. Values of F for 1 and k d.f. are the squares of t for k d.f. Here $F = 7.42$ and $t^2 = (2.73)^2 = 7.45$ which is about as close as can be expected.

When unequal numbers of observations are available in each group, the sum of squares for treatments is calculated as

$$\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} - \frac{(\Sigma X_1 + \Sigma X_2)^2}{n_1 + n_2}$$

Again note that each squared sum is divided by the number of observations in it, and that the subtracted term is the same correction factor as used in calculating the total sum of squares. For the digestibility data of section 5.7, the analysis of variance is

Table 8.3

Source of Variation	d.f.	Sum of Squares	Mean Square	F
Treatments	1	81.93	81.93	11.2**
Residual	11	80.58	7.33	
Total	12	162.51		

The value of t was 3.33; its square is 11.1.

8.3 Data with a single classification. The analysis of variance for any number of groups. In many cases, it is desired to compare more than two treatments in a single experiment. The analysis of variance is easily extended to cover such experiments. In the single classification, a number of observations are made within each group, these are at random. Thus, for example, Table 8.4 contains a record of observations made in 16 determinations of the ratio of the reacting weights of iodine and silver. There were two samples of iodine, purified by entirely different procedures, and five different preparations of silver. Determinations on eight of the ten possible iodine-silver combinations were made. These eight combinations are the treatments.

Table 8.4

Ratio of Iodine to Silver

Iodine	Silver	Ratio	Coded Ratio
I	A	1.176422	122
	A	1.176425	125
	B	1.176441	141
	B	1.176441	141
	C	1.176429	129
	C	1.176420	120
	C	1.176437	137
	D	1.176449	149
	D	1.176450	150
	E	1.176455	155
II	A	1.176399	99
	A	1.176440	140
	A	1.176418	118
	B	1.176423	123
	B	1.176413	113
	D	1.176461	161

The estimates of variance are calculated as in the previous section. The total sum of squares is first obtained, then the sum of squares attributable to treatment means; the latter is subtracted from the total to give the residual sum of squares. For the sum of squares attributable to k treatment means, calculate

$$\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \ldots + \frac{(\Sigma X_k)^2}{n_k} - \frac{(\Sigma X_1 + \Sigma X_2 + \ldots + \Sigma X_k)^2}{\Sigma n_i} .$$

This quantity has k-1 d.f. and its mean square is an estimate of σ^2. Again, note that this is not the variance of treatment means. That this is an estimate of σ^2 becomes more obvious when equal numbers of observations are made for each treatment. The sum of squares then reduces to

$$\frac{(\Sigma X_1)^2 + (\Sigma X_2)^2 + \ldots + (\Sigma X_k)^2}{n} - \frac{(\Sigma X_1 + \Sigma X_2 + \ldots + \Sigma X_k)^2}{kn}$$

which is readily seen to be (1/n)-th of the sum of squares for the variance of the treatment sums whose variance is $n\sigma^2$. Consequently, the above sum of squares needs only the divisor k-1 to be an estimate of σ^2. The above form is the generally used computation form when all n_i's are equal; i.e., sum the squares and divide by n rather than divide each squared sum by n. Note that the divisors in any case are the number of observations in the various sums.

The error sum of squares is obtained by subtracting the treatment sum of squares from the total sum of squares but can also be obtained by pooling the sums of squares of the observations within each of the k groups. The appropriate number of d.f. is the sum of the d.f. for each of the groups. For the data of Table 8.4, the analysis of variance is:

Table 8.5

Source of variation	d.f.	Sum of squares	Mean square	F
Iodine-silver combinations	7	3213.8	459	3.53 *
Residual	8	1041.7	130	
Total	15	4255.5		

As usual, the underlying distributions must be normal and have a common variance if the F-test is to be valid.

The tabled F-value at the 5% probability level for 7 and 8 d.f. is 3.50 so we conclude that the treatment differences are significant at the 5% level. This is to say that the evidence is in favor of the alternate hypothesis that differences exist. This, of course, raises the question of how the treatment combinations differ and one would naturally wonder if the differences were associated with the iodine preparations or the silver. Statistical techniques are available to answer these questions but they are a little advanced for this

chapter. In the next chapter, we will show how these questions can be answered, easily if the design of the experiment includes all ten possible iodine-silver combinations with an equal number, one or more, of observations on each combination.

While knowing that the ratios differ according to the iodine-silver combination, a further positive step has been made in treating the data by an analysis of variance procedure. When the data are considered as observations on the same population mean, the estimate of the population variance has 15 d.f. On this basis, the variance and standard deviation are 283.7 and 16.8 respectively. However, with the analysis of variance procedure, it was hypothesized as an alternative that there was more than one population mean; this possible source of additional variance was,then, eliminated from the estimate of the common variance of individuals within populations. The resulting variance and standard deviation with 8 d.f. were estimated as 130 and 11.4 respectively. This latter standard deviation is a measure of the deviation associated with multiple determinations and can be used to calculate standard errors for the means of the various combinations. This standard deviation, then, measures the reproducibility or the precision of the chemist's technique in running multiple determinations on samples of the same material.

Note that the standard deviations calculated are applicable to the coded ratios; for a standard deviation applicable to the observed ratios, decode by moving the decimal six places to the left.

8.4 Designing the new experiment. The results of the analysis of variance performed in the previous section indicate that the true ratios differ according to the iodine-silver combination used. This raised the question as to whether or not the differences could be attributed solely to the iodine sources, solely to the silver sources, to both, or to certain iodine-silver combinations. The last two possibilities are not identical.

It is immediately clear that a comparison of the two iodine sources on the basis of their means is not valid since if silvers C and E were sources of differences, they would affect the results in such a way as apparently to favor/disfavor iodine I. The same sort of argument can be used in comparing the silvers. This would not be

the case if observations were made on all possible combinations. In making observations on all possible combinations, it will be necessary to observe equal numbers to eliminate completely arguments such as above. The detection of effects due to specific combinations and not attributable to silver, iodine, or to both is also possible. This sort of experiment is a two-way classification and the subject of the next chapter.

8.5 Randomization of the variation. The sort of experiment discussed in this chapter is often referred to as the completely randomized experiment. In field experiments, the treatments are assigned to the plots at random with no restrictions on randomization. In the next chapter, the type of experiment discussed will involve restrictions on the randomization arising out of the manner in which certain major sources of variation are controlled. In other sampling experiments, a random selection is made from the individuals in each population. No requirement as to the number of observations per population is set and the experimenter may decide to take extra observations on any treatment or treatments which seem to him to merit it. Thus, for example, if the experimenter wishes to include a control treatment as a check on the experiment but does not feel that it merits as many observations as the other treatments, no real problem of arithmetic or interpretation is introduced.

The completely randomized design is used when no major source of variation is apparent in the experimental material and in need of control. For many field and laboratory experiments where a homogeneous experimental material is available, this would seem to be a very satisfactory design though in field-experiment practice, the randomized complete block design has generally proven more efficient. The measure of experimental error is seen to be the variation among plots treated alike, that is, among plots receiving the same treatment or among the observations made within each population. The mean square for treatments in the analysis of variance estimates the same population variance, under the null hypothesis. If the null hypothesis is false, an extra source of variation exists among the treatment sums or means and increases this measure of variation, on the average. It

is this increase that we are trying to detect in the usual F-test. For this reason the treatment mean square always appears in the numerator of F and significantly small values are generally considered as unusual samples or evidence of faulty assumptions or poor sampling technique rather than evidence of an incorrect null hypothesis.

8.6 The linear model; components of variance. The assumptions made about the experiment and the nature of the variation give rise to the linear model used to describe it. Thus, from the previous paragraph, the linear model states

$$X_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, \ldots, k \text{ and } j = 1, \ldots, n_i.$$

In other words, the j-th observation made on the i-th treatment or population consists of a general mean, a treatment effect τ_i, and an error ϵ_{ij}. It is assumed that the ϵ_{ij}'s are a random sample from a single population with unknown variance, and normally distributed if tests of significance are to be valid. The null hypothesis is that the τ_i's are zero. Notice that the hypothesis is easily stated when a μ is used in the model.

Treatment totals are seen to be

$$n_1\mu + n_1\tau_1 + \sum_j \epsilon_{1j}, \ \ldots, \ n_k\mu + n_k\tau_k + \sum_j \epsilon_{kj},$$

for treatments 1, ..., k. A sum of squares for these quantities could have but little meaning when the n_i's are unequal. Dividing each quantity by the appropriate n_i gives the set of means

$$\mu + \tau_1 + (\sum_j \epsilon_{1j})/n_1, \ \ldots, \ \mu + \tau_k + (\sum_j \epsilon_{kj})/n_k.$$

A variance for these quantities is seen to be more meaningful in that the $\mu + \tau_i$'s are unencumbered by multipliers; unfortunately they and the epsilon-sums are based on different numbers of observations so are measured with unequal precision. This fact is taken care of in the analysis of variance by dividing each squared sum rather than each sum, by the appropriate n. At the same time, an estimate of σ^2 is obtained from these treatment totals. It is seen, then, that the divisors do two jobs: they not only serve as divisors to assure us of an estimate of σ^2 rather than some multiple of it, but they also serve as weighting factors, weighting each mean according to the number of observations in it.

When the alternative hypothesis is true, the samples associated with the different treatments are samples from populations with different means namely $\mu + \tau_1, \ldots, \mu + \tau_k$. An estimate of the variance of these treatment means is obtained along with an estimate of σ^2 when the variance due to treatments is calculated in the analysis of variance. No such estimate is present when the null hypothesis is true.

Denote the variance of population means by σ_m^2 and define it as

$$\sigma_m^2 = \frac{\Sigma \tau_i^2}{k-1} .$$

It is clear that addition of μ to each τ_i does not affect this quantity.

In the analysis of variance with equal numbers in the different categories, calculation of the sum of squares and mean square due to treatments involves the sums $n\mu + n\tau_1 + \sum_j \epsilon_{1j}, \ldots, n\mu + n\tau_k + \sum_j \epsilon_{kj}$. Since $n\mu$ is constant, it does not affect the variance of these quantities. The variance of $n\tau_1, \ldots n\tau_k$ is $n^2 \sigma_m^2$ but the analysis of variance procedure calls for the divisor n so that the component of variance due to the τ_i's appears in the treatment mean square as $n\sigma_m^2$. Consequently, the mean square for treatments is an estimate of σ^2 and $n\sigma_m^2$.

The analysis of variance of Table 8.6 presents the ideas of the preceding paragraphs in compact form. It is seen that an estimate of σ_m^2, often denoted by $\hat{\sigma}_m^2$, is readily available. The procedure of estimation does not require the assumption of underlying normal distributions as does the F-test. Estimation of components of variance is much used by animal breeders.

Table 8.6

Analysis of Variance

Source	d.f.	S. Sqs.	M. Sq.	M. Sq. is an estimate of
Treatments	k-1	✓	✓	$\sigma^2 + n\sigma_m^2$
Within Treatments	k(n-1)	✓	✓	σ^2
Total	nk-1	✓		

The flexibility of the completely randomized design is apparent in that the number of treatments and the amount of replication on each is left to the experimenter, being limited only by the amount of experimental material available. If 30 plots are available, then two treatments can be compared with 28 d.f. in error regardless of the way in which the available replication is assigned to the two treatments. If three treatments are used, then 27 d.f. are available in error, a loss of a single degree. Even with six treatments, there are still 24 d.f. for error.

The standard deviation available from the error mean square is applicable to any of the observations. The standard error for any treatment mean is readily calculated. Since significance of F raises the problem of attributing significance to one or more treatment combinations, the idea of a quantity that can be used to test all possible comparisons has considerable appeal. Unfortunately, no such quantity is available and the experimenter is warned against testing comparisons suggested by the data themselves. Thus, when two treatment means are involved, the observed value of t exceeds the 5% level 5% of the time when the null hypothesis is true but the same quantity when used to compare the highest and lowest of three treatment means exceeds the tabulated 5% value 13% of the time when the null hypothesis is true. The percentage of times in which the comparison of the highest and lowest means exceeds the 5% tabulated value of t rises rapidly till with more than 20 treatments, significance is almost certain to be claimed even when the null hypothesis is true.

8.7 The l.s.d. A quantity often calculated when all treatments are observed an equal number of times and used inappropriately to compare all possible differences or ones that the data suggest to be of interest, is the so-called least significant difference. It is defined as that difference between treatment means that would be significant if only two treatments were involved. It is calculated as $(\sqrt{\frac{2s^2}{n}})t(.05, \text{d.f.})$ where the d.f. are those of the error mean square, s^2, and n is the number of observations per mean. This quantity is certainly useful and probably conservative for comparing adjacent ranked means arising from the data. The use of the l.s.d. in making other comparisons is to be discouraged where better methods such as those

involved in factorial experiments and where other meaningful comparisons associated with sets of d.f. can be set up. Such methods will be discussed in the next chapter. Where these are not available, the testing of differences that appear to be of interest is not to be discouraged entirely. Failure of such comparisons to attain significance is evidence that they are attributable to chance; attainment of significance is not to be considered as indicating a difference exists at the tabled probability level in use but certainly suggests the possibility of a real effect if it is well beyond such level. Other procedures now appearing show that considerable progress is being made in the matter of locating differences evidenced by a significant F.

8.8 Randomization. In a field or laboratory experiment the treatments must be assigned at random to the plots or units of experimental material. Suppose the experiment involves 3 treatments with 5, 7, and 8 replicates respectively to be assigned to 20 plots. To the plots or units, assign the number 1 to 20 in a convenient manner. From a table of random numbers draw 20 pairs of numbers in the manner of section 4.3. The first five pairs of numbers give the plots to which the first treatment is to be applied, the next 7 pairs give the plots to which the second treatment is to be applied and the remaining pairs refer to the plots to receive the third. Since the tabulated numbers run from 00 to 99, each may be divided by 5 and the remainder substituted from the tabulated pair. This involves some wastage when the number of plots is not a divisor of 100 as some numbers must be thrown away in order that the table will not present the smaller numbers with a greater frequency. Regardless of the number of plots, some number pairs will almost certainly appear more than once. When a number pair occurs for the second or later time, it is discarded.

8.9 Sampling and sub-sampling. Statistical data are often collected in a manner involving two sources of variation, both of which can be associated with error.

For example one might be sampling peas from a pea-growing area to test the significance of differences in tenderness by variety. A number of fields of each variety are selected at random and several determinations of tenderness made at each field. Some variation is expected among the subsamples within the fields. Variation is also expected among fields of the same variety. The sources of variation may be the same so that they may be comparable in magnitude but one readily admits that the within field variation may be of smaller magnitude than that between fields of the same variety.

The same is true in many other cases. Samples from all of several sources are obtained and multiple determinations, chemical or physical analyses, are made on each sample.

The two sources of error, samples and subsamples, lead to mean squares in the analysis of variance which are generally called experimental error and sampling error, respectively. Sampling error, a measure of the variance among subsamples of the primary samples, is often associated with precision as in the case of chemical analyses. In field experiments, sampling error measures something about the homogeneity of the plot material.

As an example of the analysis of such data, consider the numbers in Table 8.7. These data can be considered as observations made on the product of 3 plants in area A, 3 in B, and 2 in C. There are 14 observations, $\Sigma X = 95$, $\Sigma X^2 = 659$, and $\Sigma x^2 = 14.36$ with 13 d.f.

Table 8.7

Area	A			B			C	
Plants	I	II	III	I	II	III	I	II
Observations	6	6,8	6,7,8	5,7	6,7	6	7	7,9

This one-way classification may be first analyzed with sources of variation i) among plants ignoring areas with 7 d.f. and ii) residual. The sum of squares among plants is given by

$$6^2 + \frac{(6+8)^2}{2} + \ldots + \frac{(7+9)^2}{2} - \text{C.F.} = 5.86 \quad \text{with 7 d.f.}$$

Residual sum of squares is $14.36 - 5.86 = 8.50$ with 6 d.f.

In turn, the sum of squares among plants is attributed to areas and plants within areas. The sum of squares for areas is

$$\frac{(6+6+8+\ldots+8)^2}{6} + \ldots + \frac{(7+7+9)^2}{3} - \text{C.F.} = 4.07 \text{ with 2 d.f.;}$$

that due to plants within areas is $5.86 - 4.07 = 1.79$ with 5 d.f.

The sum of squares for both plants within areas and observations within plants may be obtained directly. For plants within areas, calculate

$$\left\{6^2 + \frac{(6+8)^2}{2} + \frac{(6+7+8)^2}{3} - \frac{(6+\ldots+8)^2}{6}\right\} + \ldots + \left\{7^2 + \frac{(7+9)^2}{2} - \frac{(7+7+9)^2}{3}\right\}$$

= 1.80 with 5 d.f. Compare 1.79 by subtraction. For observations within plants, calculate

$$\left\{6^2 + 8^2 - \frac{(6+8)^2}{2}\right\} + \ldots + \left\{7^2 + 9^2 - \frac{(7+9)^2}{2}\right\} = 8.5$$

with 0 + 1 + 2 + 1 + 1 + 0 + 0 + 1 = 6 d.f. Notice that when there is a single observation at a plant, no estimate of σ^2 is obtained and, consequently, no contribution made to the observations within plants variance.

In analysis of variance form, the results are summarized in Table 8.8.

Table 8.8

Source	d.f.	Sum of squares	Mean square
Areas	2	4.07	2.03
Plants within areas (Experimental error)	5	1.79	.36
Within plants (Sampling error)	6	8.50	1.42
Total	13	14.36	

For these "data", there is no evidence that plant to plant variation within areas is greater than the variation within plants, ($F = \frac{.36}{1.42} < 1$). In such cases and when the d.f. for experimental error are small, one very often pools the two errors to give a new estimate for testing areas. In this case, the new estimate would be

$$\frac{1.79 + 8.50}{5 + 6} = .94 \text{ with } 11 \text{ d.f.}$$

Where sufficient data are available so that there are more than ten or twenty d.f. in experimental error before pooling, pooling seems less desirable even when suggested by a test of experimental error versus sampling error. If the experimenter believes there is a source of variation in experimental error, over and above sampling error, then it is logical not to pool the two errors even when the F-test is not significant.

8.10 Variance of treatment means with sampling and subsampling.

With the possibility of both an experimental error and a sampling error, the question arises as to how to proceed with the collection of the data, in particular, whether to concentrate on getting many samples with few subsamples or determinations of each or whether to take fewer samples with more subsampling. Clearly, the answer to this will depend on the relative magnitude of experimental and sampling errors as well as upon the cost. Thus, subsamples may involve costly chemical analyses, time-consuming procedures, or destructive tests of expensive items whereas obtaining samples themselves may be of trivial difficulty. On the other hand, it may be that obtaining samples may involve expensive travel while subsampling involves little more than observation of parts of the sample. Probably, the true situation will be intermediate.

A thorough consideration of the problem of sampling versus subsampling should consider these relative costs as well as the relative magnitude of the variances involved. We shall look at the problem from the point of view of the variance of a treatment mean only.

Suppose that a manufacturer owns two plants in different parts of the country and is producing the same product, say a dairy product, at each from local materials. Since he wishes his product to be of consistent quality regardless of its origin, he decides to sample the product to test the hypothesis of a common mean. For this purpose, he draws random samples of 12 units from each dairy and runs chemical analyses on four random sub-samples of each unit. These four subsamples may be quarters but need not be.

The analysis of variance is given in Table 8.9.

Table 8.9

Source	d.f.	M. Sq.	M. Sq. is an estimate of
Factories	1		
Samples (Experimental error)	2 x 11	90	$\sigma_S^2 + 4\sigma_E^2$
Subsamples (Sampling error)	2 x 12 x 3	10	σ_S^2

Estimates of σ_S^2 and σ_E^2 are given by $\hat{\sigma}_S^2 = 10$ and $\hat{\sigma}_E^2 = (90-10)/4 = 20$. Let us consider two other possibilities, namely that 24 samples are obtained with two determinations made on each and that 8 samples are obtained with six determinations made on each. In the resulting analysis of variance, experimental error will be an estimate of $\sigma_S^2 + 2\sigma_E^2$ and $\sigma_S^2 + 6\sigma_E^2$ respectively. Foom the "data" of Table 8.9, we estimate these as $10 + 2 \times 20 = 50$ and $10 + 6 \times 20 = 130$ for the two schemes.

Factories are to be compared with variation among samples as the basis for judging differences in the quality of the factory product. Hence the comparison is between factory means which contain $24 \times 2 = 48$ and $8 \times 6 = 48$ observations in each case. The variance of treatment mean is given by 50/48 and 130/48 according to the sampling scheme used, one variance being two to three times that of the other.

As a measure of the relative efficiency of the two methods, one might take the ratio of the variances of a treatment mean. Here, we have $(130/48) \div (50/48) = 2.6$. We conclude that the scheme with the two determinations is 260% as efficient as that with six determinations. The gain in efficiency is 160%.

Other schemes may be considered where the number of observations per factory is not 48. The question of whether or not to take subsamples must be up to the experimenter. He will often wish them to measure the precision of a physical or chemical procedure at the expense of experimental error and the variance of a treatment mean.

CHAPTER 10
LINEAR REGRESSION

10.1 Summary. This chapter deals with the calculation and use of a linear prediction equation. Estimation of confidence limits for predicted values and for the regression coefficient are treated. Tests of hypotheses are carried out for the regression coefficient and for the equality of two regression coefficients.

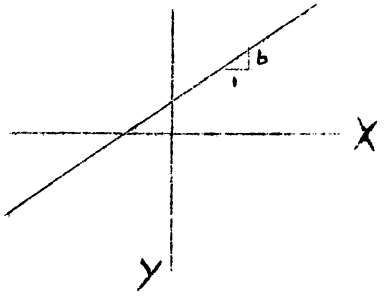
10.2 Introduction. Up to this point, we have been concerned mainly with a single variable and the manner in which it varies. Only briefly have we referred to concomitant measurements and joint variation. It is to be noted that concomitant observations were made when note was taken of the replicate in which a certain plot appeared and the treatment that was applied to that plot. However, we were rarely concerned with a measureable concomitant observation except where we dealt with equally spaced treatments.

There are many situations where a pair of observations is obtained. In some cases where both observations involve a measurement, we notice that there is a more or less well-defined relationship between the variables. For example, in adults, large values of weight appear to be associated with large values of height, low weights with low heights; yield of wheat is depressed as the degree of infection with stem rust increases; achievement in class as measured by numerical grades increases with ability as measured by the score obtained on an IQ examination; many other examples will occur to the reader.

In some of the cases, a relationship will be very pronounced while in other cases, no relationship may be apparent. Where a relation is apparent, a cause and effect system may be obvious as in the case of wheat yield and stem rust infection; on the other hand, only a joint variation as in the case of adult heights and weights may be apparent. There may be theories about existing relations or simply the observation that they exist. In any case, we will wish to make use of such relations. For most of us, this will consist of predicting one variable from another and for this purpose, we determine a mathematical relationship such as in chapter 3 for determining probabilities. In this chapter, we shall be interested only in straight line or linear relations, i.e. those expressible in the form

$Y = a + bX$, where a and b are known constants, X is a given value of one variable and Y is the predicted value of the second variable. When $X = 0$, $Y = a$; this point is called the X-intercept of the line. When $a = 0$, the line goes through the origin. The number b is called the slope of the line and measures the increase in Y per unit of X. (See figure 10.1) When b is positive, the line slopes from the lower left to the upper right portion of the graph; when b is negative, the line slopes from upper left to lower right. With both the slope of the line and the Y-intercept, it is seen that the position of the line in the diagram is uniquely determined.

Figure 10.1



10.3 A linear relation and a prediction. Consider the data of table 10.1. It is desired to predict the number of horses for 1949 on the assumption that the decrease in the number of horses is linear with respect to year.

Table 10.1

Year	Number of horses on Canadian farms
1944	2735
1945	2585
1946	2200
1947	2032
1948	1904

To do this, let X equal the year variable and Y the variable for the number of horses. Calculate the quantities Σx^2 and Σy^2 as in previous chapters and the new quantity Σxy defined as follows:

$$\Sigma xy = \Sigma(X-\bar{x})(Y-\bar{y}) = \Sigma X(Y-\bar{y}) = \Sigma(X-\bar{x})Y$$

The calculation formula is:

$$\Sigma xy = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n} .$$

The similarity between this working formula and that for Σx^2 will be noticed if Y is replaced by X in the calculation formula. For the

horse-year data, values of these quantities are:

$$\Sigma x^2 = 10,\ \Sigma xy = -2{,}215,\ \Sigma y^2 = 508{,}702.8.$$

To calculate Σx^2, a convenient coding would be to subtract 1940 from each X, leaving the coded values of 4, 5, ..., 8. These numbers can also be used in calculating Σxy. No decoding is required. The quantities a and b are calculated as:

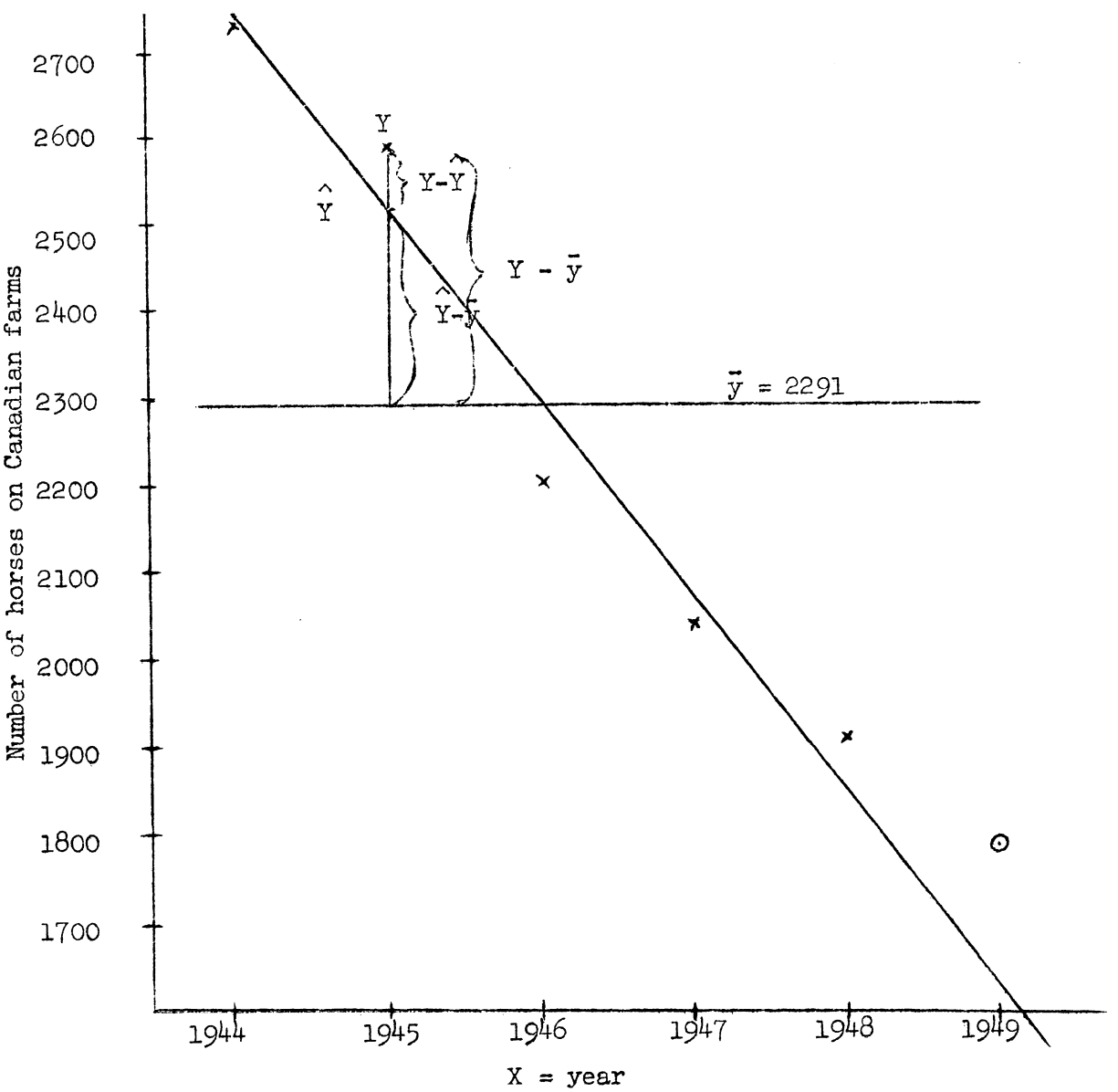
$$b = \frac{(\Sigma xy)}{(\Sigma x^2)} = -221.5 \quad \text{and} \quad a = \bar{y} - b\bar{x} = 433{,}330.2$$

From the nature of a, it is readily seen that the linear equation may be written in the form $Y - \bar{y} = b(X - \bar{x})$, informative if not compact. Our linear prediction equation is

$$Y = -221.5\,X + 433{,}330.2.$$

This is graphed, together with the paired observations constituting the data, in figure 10.2.

Figure 10.2



To use our equation to predict the number of horsee for 1949, substitute 1949 for X and solve for Y to obtain 1627. The number of horses for 1949 was 1796. Our prediction equation doesn't seem to have done so very well for us. This suggests the need of some measure of the reliability of our prediction.

Let us consider our equation a little more closely by "predicting" values of Y for each of the years 1944 through 1948. Since these are predicted values, denote tham by $\hat{Y}$ to distinguish them from the observed Y values. Both are given in table 10.2. The "predicted" values can be used to construct figure 10.2, although any two pairs of values are sufficient. The table also contains the signed differences $Y - \hat{Y}$, between the observed and predicted values of Y. Note that $\Sigma(Y-\hat{Y}) = 0$ and none of the predicted values equals its corresponding observed value. However, the differences are not greater than 5% of the mean Y. For 1949, $Y - \hat{Y} = 169$. For 1943, $\hat{Y}$ is 2956; Y is 2775; and $Y - \hat{Y}$ is -181. In extrapolation, there has been a considerable error. We were prepared to accept a straight line or linear relationship for 1944 through 1948 and assumed that it was linear back to 1943 and would continue linear into 1949. Here there is a warning against assuming the line can be continued into either the past or the future as a straight line. Extrapolation beyond the range of the X values is often subject to such risks. On the other hand, a prediction that proves unreasonable may simply be that unusual case that happens once in 20 times if that is the probability assigned to type I errors. Finally, some function of the $Y-\hat{Y}$ values is indicated as a measure of judging the reliability of a prediction.

Table 10.2

X = year	Y = No. horses	$\hat{Y}$ = Predicted no. horses	$Y - \hat{Y}$
1944	2735	2734	1
1945	2585	2513	72
1946	2200	2291	-91
1947	2032	2070	-38
1948	1904	1848	56

Since the prediction of a Y value for an X within the range of observed X's is not especially informative for other than the year values, and since we are not in the business of deciding that the trend will or will not continue to be linear, our example is not a very useful one. However, it is easy to find useful examples and

we will deal with at least one in this chapter. Finally, there may be evidence or a theory for assuming that linear extrapolation is valid under some conditions for certain problems.

10.4 Concerning prediction. This example of prediction is not our first. The discussion of estimation of the mean μ and of confidence limits in chapter 2 involved prediction. In such cases, apparently a mean value was predicted for it is difficult to see what prediction of an individual value could mean. This could well have been only because of the examples chosen, for consider the following example:

The use and development of water has grown rapidly. As a consequence, there is a serious demand for advance estimates of the volume of water or the rate of flow as supplied by streams in a watershed. Thus, hydrologic forecasting is desirable. Some of the problems are immediately obvious. The amount of snowfall is an obvoous variable to measure. Careful thought by anyone who has shovelled snow suggests that snow may be variable in its water-content; it may fall on rocky ground where all must run off or on soil and soil-cover capable of holding large amounts of moisture; the local climate may be customarily clear and dry and the air capable of picking up much surface-moisture, or it may be customarily cloudy and dull to the point of being foggy. The choice of a factor that can be measured in advance of run-off and yet be highly associated with it, in the sense of their varying together, does not appear to be a simple one.

An obvious approach would be to calculate the mean run-off for a number of preceding years and make this mean our prediction. It will be seen that in the case of our linear prediction equation, a predicted value is a mean also, though we may speak as though we are predicting an individual value. The "individual value" part of the prediction is introduced by means of an appropriate standard error, our measure of accuracy. The new X will be independent of $\bar{x}$ as calculated from the previous observations so that $X - \bar{x} = -(\bar{x}-\mu) + (X-\mu)$ and the deviation of the new X from the old mean, $\bar{x}$, is seen to have a variance $\sigma^2/n + \sigma^2$ which is estimated as $s^2(\frac{1}{n} + 1)$. When we have completed our calculations, involving Student's "t", we will be able to make the following sort of a forecast: "Next year's run-off will be in the stated interval unless the run-off is so unusual as to occur less than once in 20 years in the long run",

where for the number 20 the individual may substitute the number of his choice provided he makes appropriate adjustment in his two values. It has been assumed that the sample of X-values is drawn at random from a normal population. Some work has been done on the choice of a concomitant variable for water run-off but we will not concern ourselves with this problem.

The mean number of horses on Canadian farms between 1944 and 1948 inclusive was 2291. With no knowledge of the year, this would have to be our predicted value for any year. Clearly the value predicted from our maligned equation, namely 1627, is of more intrinsic value than 2291. This will become more apparent when we see how to obtain a measure of accuracy for the value predicted from the linear equation. This will, of course, be on the assumption that the relationship between X and Y is and continues to be linear.

10.5 The linear model. The regression of Y on X is defined as the mean value of a variate Y in a distribution where the values of X are fixed, when the mean value of Y is regarded as a function of the fixed variate X.

Regression is, then, covered by a definition. Let us look carefully at it.

For any experimental design, a model is chosen depending on the design. This model defines a mean for each cell as the sum of certain effects. The observation is made on the mean but contains an additional random element denoted by ϵ. The make-up of the mean permitted us to estimate its components (which wasn't done) and still obtain an estimate of the variance of the ϵ's.

The definition of regression is very similar. It states that a mean of a variate Y is a function of an associated X rather than of an associated cell. Thus, if all possible values of Y were available for a single X, the mean μ_y would be a function of X. For linear regression, the mean could be written as

$$\mu_y = \alpha + \beta X,$$

a constant α plus a multiple or fraction β, of the X involved. β may be positive or negative. The definition also states that if values of Y were obtained for a different X, the mean of this new set of values would be of the same form as the old; i.e., α and β would not change, only X. In general, then, if α and β are known, it is only

necessary to observe an X or state anX of interest, and the mean of Y for this X is calculable.

When an observation is made, it will consist of a pair of values, (X_i, Y_i), the X_i being a matter of choice if desired or being left to chance. The value Y_i will be an observation on the mean but will have an associated random error. The model is

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

The primary observation is Y_i, the concomitant observation X_i is associated with the population mean as were the block and treatment in the randomized complete block design.

The problem is to estimate α and β and the variance of the ϵ's. A sample of (X, Y) values is available in which the Y's, at least, are random. These values give a line

$$Y = a + bX$$

where a and b are the estimates of α and β, which can be used to predict or estimate a value of Y for an X that may be of interest and an estimate of the variance of the ϵ's which can be used in assigning a measure of reliability to our prediction or estimate. It is important to note that our prediction or estimate is of a mean rather than an observation as was pointed out in section 10.4.

As far as regression is concerned, it is to be noted that the X's are observable parameters whereas the parameters α, β and σ^2 can only be estimated. We are dealing with a family of populations, one population for each X, rather than a single population. The mean value of the variate Y is expressed as a function of the fixed variate X, now a parameter associated with the individual population. Thus, a regression equation is seen to supply us with a moving average.

At this stage, it is worthwhile pointing out that we are dealing with statistical laws, laws that hold on the average. In general, the lines which we estimate in problems of regression are lines about which the observed pairs of values cluster and not lines on which the points are expected to fall. The statistical relation Y = a + bX estimates a mean and, consequently, determines a frequency distribution when a value of X is substituted.

A functional relation assigns a value rather than a distribution. For this sort of a relation, see section 3.3 where probabilities were determined from equations or section 3.6 where a functional relation assigns an ordinate in a normal curve.

In the example of the beginning of this chapter, the fixed variates or known parameters were the year values; the random variates were the numbers of horses. Only one value of Y was obtainable and we assume it to be a random value from all possible values there might have been. The problem is to estimate the mean value for each year on the assumption, not a part of our definition, that the regression of Y on X exists and is linear.

In the regression equation, that is, the mean value of Y as a function of the fixed variate X, substitute any desired value of X, perform the indicated calculation, and the result is our prediction of Y for the chosen X value. By definition, our predicted value is a mean. Often, this is what we will wish to predict. At other times, as in the example of section 10.4, we will wish to predict an individual value. In this case, use the value given by the regression equation but attach to it a variance appropriate for individuals. Thus, we will often speak of a predicted value as though it were not a mean, and this is seen to be a reasonable procedure.

The value of X is to be measured without error, i.e. there is to be no error of observation. This is important in that for the calculation of a variance, we have a sample of Y-values (it may consist of a single Y) for a given X-value. Thus if the Y-values include values for the given X together with values of Y for X's near the given one and measured in error as the given one, then we have Y values from populations other than the desired one, and this will almost certainly lead to errors in the variance if not in the mean. The variate X does not have a variance, as we have been using the term, in a regression equation although the sum of squares associated with it is usually calculated as we did in our earlier example. All of this is not to say that we cannot treat statistically problems where X is subject to error but simply that we will not in these notes.

The fixed variate is often called the <u>independent variable</u> or the <u>argument</u> and the one whose value is determined subject to random sampling error according to the value of the independent variable is called the <u>dependent variable.</u> These terms are in general use.

A measure of reliability to be associated with our prediction was not discussed in the example. Consider now an example where a number of observations are made at each of a few values of the independent variable. With this example, we shall discuss measures of

reliability and the assumptions that are required if we are to make probability statements about our predictions. Data of this type lead to a good example as far as explaining regression is concerned but are generally treated by simpler arithmetic. Here, the arithmetic is carried out for the general case where the observations occur in pairs and none of the X's need to be equal.

10.6 Prediction and reliability. An experiment with rice was carried out to consider the combined effect of number of plants per hill and distance between hills. Five values for plants per hill were used and 6 values of distance between hills. The experiment was replicated 10 times. Let us consider the data for 7 plants per hill only and use only three distances bewween hills; this is merely to reduce the calculations in our example. We have chosen the distance values of 6, 12, and 18 inches for further ease of computation in some future analysis of variance comparisons we plan. These data are presented in table 10.3 and in figure 10.3.

The calculations proceed as before:

$\Sigma x^2 = 720$ (in)^2, $\Sigma xy = -5{,}346$, $\Sigma y^2 = 122{,}173.47$ (gms)^2, $\bar{x} = 12$ in., $\bar{y} = 352.13$ gms., $b = 7.425$ gms/in. between rows.

Note how easily Σx^2 can be calculated with equally-spaced values of X and repetition. It can be shown that the constant a is of the form $\bar{y} - b\bar{x}$. Consequently the prediction equation may be written in the form

$$Y - \bar{y} = b(X - \bar{x})$$

often more convenient for many users since X-and Y-values are in the form of deviations from the mean and at the same time easily used for the calculation of Y-values by simply transposing $\bar{y}$. Also we see the composition of a instead of the mixture of ingredients. In this form, our equation becomes

$$\hat{Y} - 352 = 7.4(X - 12)$$

The hat on Y is to point out that an estimate of Y, rather than an observed value, is involved.

The criterion which leads to the equation given above is that the sum of squares of the distances between the observed points and the corresponding mean given by the line $Y = a + bX$ be a minimum. This is equivalent to dropping lines perpendicular to the X-axis from the observed points to the proposed line. Of all such possible lines, the one chosen is associated with a minimum sum of squares.

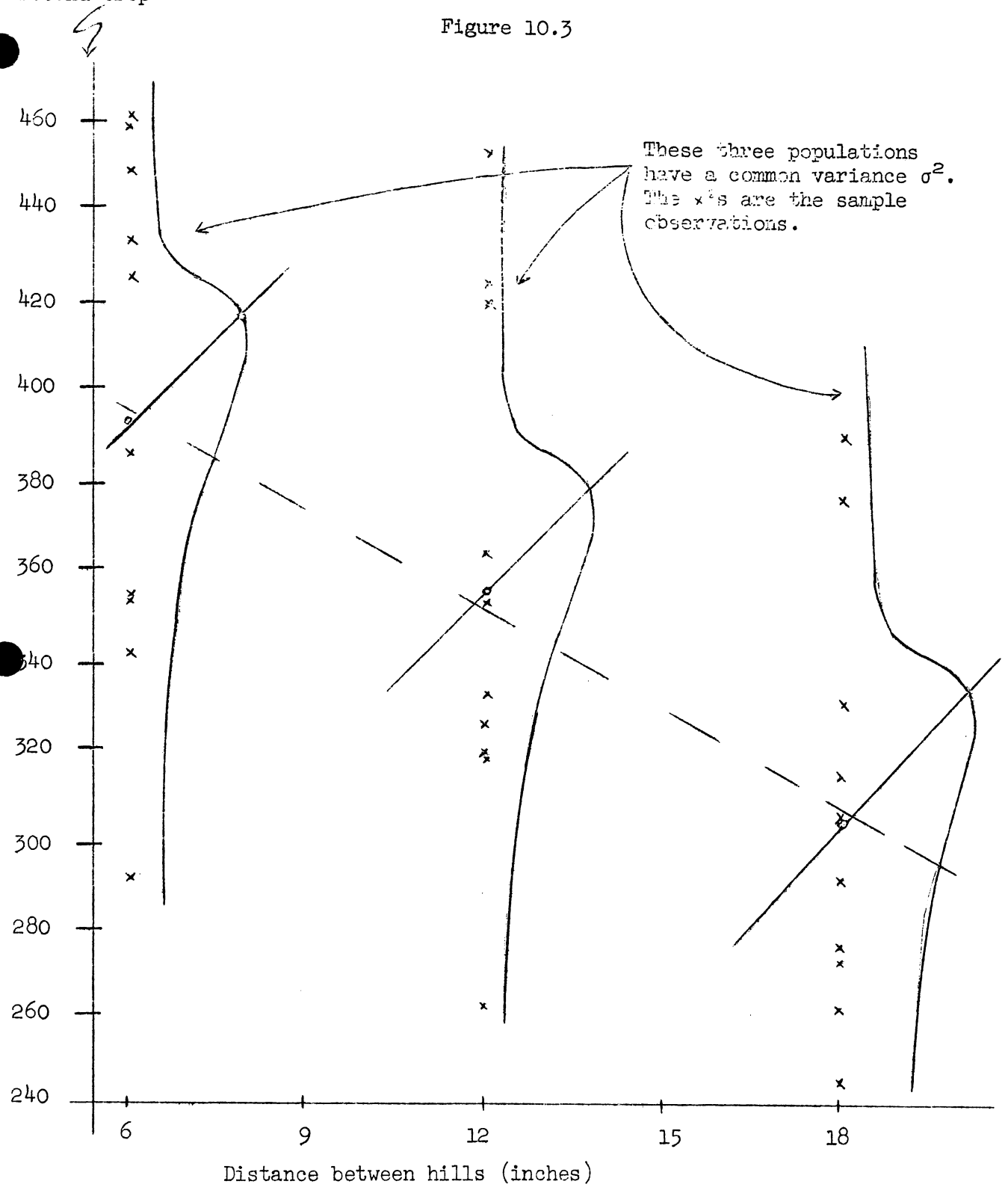
Geometrically, this is a bit awkward to depict in figure 10.3 as our 30 points lie opposite only 3 X-values. However, you can readily see what is involved by looking at figure 10.3 where all five distances can be seen. Algebraically, consider all possible straight lines, say $\tilde{Y} = a + bX$, and select that one for which $\Sigma(Y-\tilde{Y})^2$, where Y is an observed value and $\tilde{Y}$ is the corresponding value from the prediction equation, is a minimum. The $\tilde{Y}$ chosen by this procedure is denoted by $\hat{Y}$ and $\Sigma(Y-\hat{Y})^2$ is smaller than the sum of squares for any other choice of straight line. No assumptions concerning the individual paired values are required by the procedure. Assumptions are required when probability statements are to be made.

Table 10.3

Distance between hills (inches)	Yield in grams, second crop	
6	423	
	433	
	460	
	386	
	458	
	354	
	290	
	354	
	341	
	447	
		3946
12	453	
	364	
	418	
	353	
	423	
	317	
	325	
	262	
	331	
	317	
		3563
18	376	
	329	
	306	
	314	
	388	
	260	
	245	
	276	
	271	
	290	
		3055

(7 plants per hill data) Jour. Amer. Soc. Agronomy Vol. 29, 1937 pp. 181

Figure 10.3


Yield in grams, second crop
460
440
420
400
380
360
340
320
300
280
260
240
These three populations have a common variance σ^2. The x's are the sample observations.
6
9
12
15
18
Distance between hills (inches)

This criterion is not the only one possible for choosing a straight line. We have used it before so that we are being consistent. If you refer to section 2.10, you will see a proof that $\Sigma(X-\bar{x})^2$ cannot be made smaller by replacing $\bar{x}$ by any other number. In the experimental designs discussed, $\bar{x}$, an estimate of μ, was replaced by a more complicated value which we did not estimate but which depends on the linear model and varies from cell to cell. Regression is similar to experimental design in that a moving average, one that varies according to the value of X, is called for. Then, on the basis of the moving average, exactly the same principle as for the fixed average was used. The principle is called the Principle of Least Squares. Similarly, $\Sigma(Y-\hat{Y}) = 0$.

In section 2.7, it was shown that $\Sigma(X-\bar{x}) = 0$./ Consequently, $\Sigma Y = \Sigma\hat{Y}$ where the $\hat{Y}$'s consist only of those values, but all of them, which correspond to observed Y's. You can demonstrate this for the simple example given in table 10.2. It is easy to show algebraically for the general case. The quantities $Y-\hat{Y}$ are called, as before, deviations or residuals.

Another property of our line is that it passes through the general mean $(\bar{x},\bar{y})$. To show this, replace X in the general equation

$$\hat{Y}-\bar{y} = b(X-\bar{x})$$

by $\bar{x}$. Then $\hat{Y}-\bar{y} = 0$, i.e. $\hat{Y} = \bar{y}$ for $X = \bar{x}$ and consequently $(\bar{x},\bar{y})$ is a point on the straight line.

Now consider the sum of squares of residuals. This is

$$\Sigma(Y-\hat{Y})^2 = \Sigma\left[(Y-\bar{y}) - b(X-\bar{x})\right]^2$$

Notice that the square is outside the outermost bracket. From this

$$\Sigma(Y-\hat{Y})^2 = \Sigma\left[(Y-\bar{y})^2 - 2b(Y-\bar{y})(X-\bar{x}) + b^2(X-\bar{x})^2\right]$$

$$= \Sigma(Y-\bar{y})^2 - 2b\Sigma(Y-\bar{y})(X-\bar{x}) + b^2\Sigma(X-\bar{x})^2$$

$$= \Sigma(Y-\bar{y})^2 - \frac{2\left[\Sigma(Y-\bar{y})(X-\bar{x})\right]^2}{\Sigma(X-\bar{x})^2} + \frac{\left[\Sigma(Y-\bar{y})(X-\bar{x})\right]^2}{\left[\Sigma(X-\bar{x})^2\right]^2}\Sigma(X-\bar{x})^2$$

$$= \Sigma(Y-\bar{y})^2 - \frac{\left[\Sigma(Y-\bar{y})(X-\bar{x})\right]^2}{\Sigma(X-\bar{x})^2}$$

In the second part of this expression notice that the numerator is the square of a sum while the denominator is a sum of squares. The first term is simply the variance of the Y's so that the second part is the reduction due to regression, i.e. due to choosing an average dependent on X. The sum of squares about a regression line or moving average is similar to that used in dealing with the fixed mean for a sample of observations from a single population or the cell means of the various experimental designs. It is the numerator of an estimate of the variance of the ϵ's in our model and is used as a measure of accuracy in making predictions. In this case, the divisor is n-2, the number of degrees of freedom. The loss of an extra degree of freedom is associated with the estimate of β, namely b, and in turn with the reduction due to regression. This reduction has been seen to be $\left[\Sigma(X-\bar{x})(Y-\bar{y})\right]^2/\Sigma(X-\bar{x})^2$. The mean square, $\Sigma(Y-\hat{Y})^2/(n-2)$, is called the variance about regression and is denoted by $s^2_{y.x}$; its square root is called the standard deviation from regression or the standard error of estimate.

Consider the partitioning of the sum of squares of the Y-values. Since $\Sigma(Y-\hat{Y})^2 = \Sigma(Y-\bar{y})^2 - \left[\Sigma(X-\bar{x})(Y-\bar{y})\right]^2/\Sigma(X-\bar{x})^2$, it follows that

$$\Sigma Y^2 = \frac{(\Sigma Y)^2}{n} + \frac{\left[\Sigma(X-\bar{x})(Y-\bar{y})\right]^2}{\Sigma(X-\bar{x})^2} + \Sigma(Y-\hat{Y})^2 .$$

This can be related to the equation

$$Y = \bar{y} + (\hat{Y}-\bar{y}) + (Y-\hat{Y})$$

whose validity can be seen by simply removing the brackets. This relation can be seen easily in figure 10.4 for the year 1945. The relationship between the partitioning of Y and the partitioning of the sum of squares is readily seen. The first parts are the mean and the correction factor for the mean, respectively. That the second parts are related is seen from the following simple algebra:

$$\frac{\left[\Sigma(X-\bar{x})(Y-\bar{y})\right]^2}{\Sigma(X-\bar{x})^2} = \left(\frac{\left[\Sigma(X-\bar{x})(Y-\bar{y})\right]^2}{\Sigma(X-\bar{x})^2}\right)\left(\frac{\Sigma(X-\bar{x})^2}{\Sigma(X-\bar{x})^2}\right)$$

$$= b^2\Sigma(X-\bar{x})^2$$

$$= \Sigma\left[b(X-\bar{x})\right]^2$$

Recall that $(\hat{Y}-\bar{y}) = b(X-\bar{x})$. The correspondance is now clear. Finally, the last parts of the two equations are clearly comparable.

The results may be presented in an analysis of variance table such as table 10.4. The numbers are for the rice data.

Table 10.4

Source	d.f.	Sum of Squares (Algebraic)	Sum of Squares (Rice data)	Mean Square
Reduction	1	$\frac{[\Sigma(X-\bar{x})(Y-\bar{y})]^2}{\Sigma(X-\bar{x})^2}$	39,694	39,694
Residual	n-2=28	Difference	82,479.47	2,945.70
Total	n-1=29	$\Sigma(Y-\bar{y})^2$	122,173.47	

The regression equation is repeated:

$$\hat{Y} = 352 + 7.4(X-12)$$
$$= 263.2 + 7.4X \ .$$

In many texts, you will not find a hat on the Y in the equation. After you have gained some familiarity with regression equations, this will not be confusing.

Up to this point, no assumptions about our pairs of values have been necessary. In the analysis of variance table, a sum of squares has been partitioned by a purely algebraic process. One of these parts measures the variation about the regression line and is an obvious candidate for use as a measure of reliability, either in the setting of confidence limits or the testing of hypotheses; the other part can serve as a measure of the worth of the assumption of linear regression just as replicate or treatment means squares serve that purpose in analysis of variance.

10.7 Assumptions and probability statements. To make any exact probability statements, randomness of the Y-values is necessary. For normal theory to be applicable it is assumed that the deviations about regression are random and normally distributed with a common variance. In terms of the linear model, the ϵ's must be a random sample from a single normal distribution. Thus, randomness is seen to refer only to Y-values. Our estimate of σ^2, the variance of the ϵ's, is given by $\Sigma(Y-\hat{Y})^2/(n-2)$, i.e. the mean square of the deviations from regression, and is an unbiased estimate of σ^2. It is denoted by $s^2_{y.x}$.

In the first simple prediction of a future value (section 10.4), the variance of the sample mean as well as that of the variable was used in assigning a measure of reliance to our prediction. The same is true for a prediction based on a moving average, a regression line. In this prediction, the mean involves not only the mean of the Y-values but also a regression coefficient and an $(X-\bar{x})$-value. The value of $(X-\bar{x})$ is a constant chosen by the person making the prediction but the number b is a variable. Since b is a variable, it has a distribution with a mean and variance and some knowledge of this variance is required to assign a measure of reliance to our prediction. From the computing form of b, it is seen to be a linear combination of the Y's; i.e., the Y's are raised only to the first power, there are no products of Y's, and their multipliers are constants, values of $X-\bar{x}$. This makes the variance of b easy to calculate when the Y's are normally distributed. The variance of b is

$$\frac{s^2_{y\cdot x}}{\Sigma(X-\bar{x})^2}$$

For the rice example, the variance of b is

$$2{,}945.70/720 = 4.09125$$

and the standard deviation is 2.02 gms/in.

A glance at figure 10.3 indicates why the variance of b must be considered in obtaining a measure of reliability for our prediction. The mean $(\bar{x},\bar{y})$ of the observed pairs of observations may lie either above or below the true value $(\bar{x},\mu_y)$. At the same time, the slope, b, of the line may be either greater or less than the true value β. Thus a prediction has two sources of error, each of which has a variance. Note from the figure that an error in the estimate of β causes more trouble in the predicted Y-value according as the chosen X is farther and farther from $\bar{x}$. Thus, the quantity $(X-\bar{x})$ which appears in our prediction should also appear in the variance of our prediction.

A new yield, then, will vary from the sample average, i.e. the predicted value, as provided by the regression equation. This variation is expressed by the following equation:

$$Y - \left[\bar{y} + b(X-\bar{x})\right] = \left\{Y - \left[\mu_y + \beta(X-\bar{x})\right]\right\} - (\bar{y}-\mu_y) - (X-\bar{x})(b-\beta)$$

(Remove brackets to check this equality.)

The variances of the three components are known. The first variance is that of individuals about the population regression line for X equal to the desired value; the second is that of means about the population regression line for $X = \bar{x}$; and the third is a constant multiple, $X-\bar{x}$, of that of the b-values. The resulting variance of the prediction of a future Y-value is

$$\sigma_{y.x}^2 + \frac{\sigma_{y.x}^2}{n} + \sigma_{y.x}^2 \left[\frac{(X-\bar{x})^2}{\Sigma(X-\bar{x})^2}\right] = \sigma_{y.x}^2\left(1 + \frac{1}{n} + \frac{(X-\bar{x})^2}{\Sigma(X-\bar{x})^2}\right)$$

An estimate of this variance is obtained by substituting $s_{y.x}^2$ for $\sigma_{y.x}^2$.

Finally, we make a probability statement about the future Y-value. Student's "t" is required and the statement is that the future Y-value for the specified X will lie between

$$\bar{y} + b(X-\bar{x}) - t(.05, n-2)\, s_{y.x}\left[1 + \frac{1}{n} + \frac{(X-\bar{x})^2}{\Sigma(X-\bar{x})^2}\right]$$

and

$$\bar{y} + b(X-\bar{x}) + t(.05, n-2)\, s_{y.x}\left[1 + \frac{1}{n} + \frac{(X-\bar{x})^2}{\Sigma(X-\bar{x})^2}\right],$$

unless a chance as unlikely to occur as one time in 20, does occur. The probability level may be chosen as desired. For convenience, this is written as

$$P\left\{\bar{y} + b(X-\bar{x}) - t(.05, n-2)s_{y.x}\left[1 + \frac{1}{n} + \frac{(X-\bar{x})^2}{\Sigma(X-\bar{x})^2}\right]\right.$$

$$\left. \leq \hat{Y} \leq \bar{y} + b(X-\bar{x}) - t(.05, n-2)s_{y.x}\left[1 + \frac{1}{n} + \frac{(X-\bar{x})^2}{\Sigma(X-\bar{x})^2}\right]\right\}$$

$$= .95,$$

or

$$P\left\{\hat{Y} - t(.05, n-2)s_{y.x}\left[1 + \frac{1}{n} + \frac{(X-\bar{x})^2}{\Sigma(X-\bar{x})^2}\right]\right.$$

$$\left. \leq Y \leq \hat{Y} - t(.05, n-2)s_{y.x}\left[1 + \frac{1}{n} + \frac{(X-\bar{x})^2}{\Sigma(X-\bar{x})^2}\right]\right\}$$

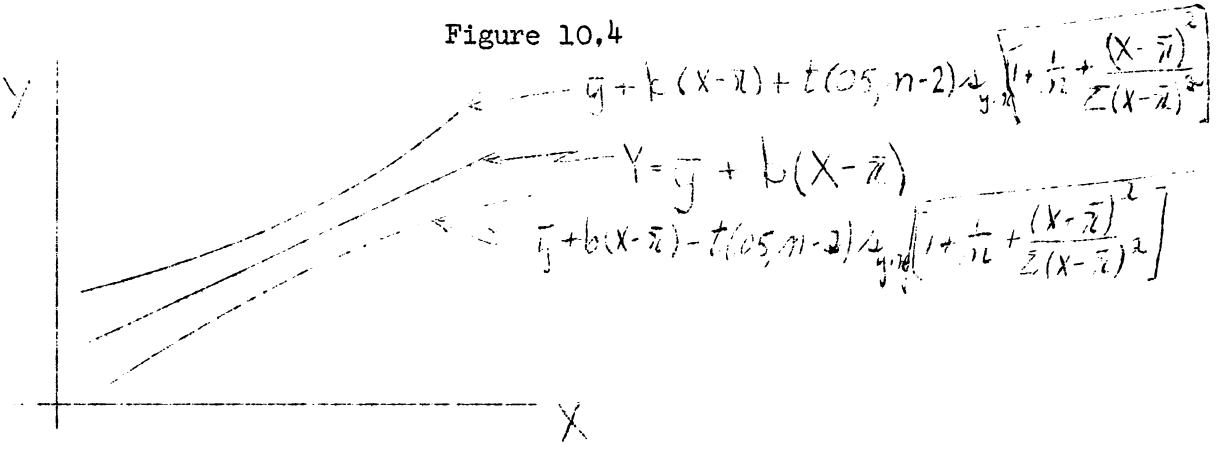
$$= .95$$

and is not interpreted as an exact probability statement since the new Y will either fall within the interval or it won't. The probability applies not simply to the next Y to be observed but to the whole

event consisting of the collection of the data which led to the prediction equation and future observed Y.

Because of the $(X-\bar{x})^2$ in our confidence limits, (under the square root sign and a part of $s_{y.x}$), the length of the interval is not constant for all X. Instead, there is a confidence band as shown in figure 10.4.

Figure 10.4



The confidence interval is exact for a single prediction. Repeated prediction using the same equation affects the probability associated with the usual statements and is not recommended.

10.8 Estimation of and confidence limits for means. Here we have the usual problem of estimation, estimation of the mean of a population, the population of Y's associated with a specified X. The estimate will be the same value used for a prediction but the variance will be appropriate for a mean rather than an individual. Here, a value for a parameter rather than a value of a future observation is required. Denote this parameter by $\mu_{y.x}$, the population value of the moving average of Y for the stated value of X. The estimate is supplied by the prediction equation derived from the sample. Thus,

$$\hat{\mu}_{y.x} = \hat{Y} = \bar{y} + b(X-\bar{x}).$$

In assigning a measure of reliability to our estimate, consider the deviation of the estimate from the true value, viz.

$$\hat{Y} - \mu_{y.x} = \bar{y} + b(X-\bar{x}) - \mu_{y.x}$$

rather than the deviation $Y-\hat{Y}$ of a future random sample from an estimated mean. The variance here involves that of a sample mean based on n observations and a fixed set of X's, n=30 for the rice data, the variance of b and our choice of an X-value. The quantity

$\mu_{y.x}$ is a location parameter and has no variance. The required variance is

$$\frac{\sigma^2_{y.x}}{n} + \sigma^2_{y.x}\left[\frac{(X-\bar{x})^2}{\Sigma(X-\bar{x})^2}\right] = \sigma^2_{y.x}\left[\frac{1}{n} + \frac{(X-\bar{x})^2}{\Sigma(X-\bar{x})^2}\right].$$

Notice this differs from the variance for an individual by the absence of an additional one times the variance.

Confidence limits are given by use of Student's t in the usual manner. Thus, using a 95% level of confidence we say that the population mean, $\mu_{y.x}$, lies between

$$\bar{y} + b(X-\bar{x}) - t(.05,\ n-2)s_{y.x}\sqrt{\frac{1}{n} + \frac{(X-\bar{x})^2}{\Sigma(X-\bar{x})^2}}$$

and

$$\bar{y} + b(X-\bar{x}) + t(.05,\ n-2)\ s_{y.x}\sqrt{\frac{1}{n} + \frac{(X-\bar{x})^2}{\Sigma(X-\bar{x})^2}}$$

unless a chance as unlikely to occur as one time in 20 has occurred.

In the special case where the X of interest is chosen as $\bar{x}$, $X - \bar{x} = 0$, and the estimate becomes $\bar{y}$ with a variance $s^2_{y.x}/n$. The confidence interval is now given by the two quantities

$$\bar{y} \pm t(.05,\ n-2)\ s_{y.x}\sqrt{\frac{1}{n}}.$$

Thss is not surprising since the regression line goes through the point $(\bar{x},\bar{y})$ and changing b does not affect $\bar{y}$, the estimate of $\mu_{y.x}$. Consequently the variance of b should not affect the reliability of the estimate of $\mu_{y.x}$ when $X = \bar{x}$.

10.9 Estimation of and confidence limits for the regression coefficient. An unbiased estimate of β, the population regression coefficient, is given by

$$b = \frac{\Sigma(X-\bar{x})(Y-\bar{y})}{\Sigma(X-\bar{x})^2}.$$

Since it is a linear combination of normally distributed Y's, its variance is calculated as

$$\frac{\sigma^2_{y.x}}{\Sigma(X-\bar{x})^2}$$

and estimated by

$$\frac{s^2_{y.x}}{\Sigma(X-\bar{x})^2}$$

For confidence limits on the parameter β, Student's "t" is used, the confidence limits being

$$b \pm t(.05,\ n-2)\sqrt{\frac{s^2_{y.x}}{\Sigma(X-\bar{x})^2}}\ .$$

The statement that

$$b - t(.05,\ n-2)\sqrt{\frac{s^2_{y.x}}{\Sigma(X-\bar{x})^2}} \leq \beta \leq b + t(.05,\ n-2)\sqrt{\frac{s^2_{y.x}}{\Sigma(X-\bar{x})^2}}$$

will be correct unless a chance as unlikely to occur as one time in 20 has occurred.

10.10 Tests of hypotheses. Point estimation, i.e. a single value, and interval estimation, i.e. confidence limits, have been discussed. Often, it is desired to perform a test of hypothesis first to determine whether or not to proceed with the estimation. While confidence limits can be used for testing hypotheses, the computations required for the usual tests of hypotheses are sometimes less time-consuming. In some cases, the test of hypothesis is considered as the end result prior to the drawing of conclusions.

An obvious hypothesis to test is whether it has been worth while to use a prediction equation. It has been worth while if the regression line has a slope so different from zero that it is difficult to explain on a chance basis or if the reduction in the variance of Y due to using a regression line is too large to be readily attributable to chance. The implied questions turn out to be the same and the answers are supplied by the test criteria t and F respectively where $t^2 = F$.

To test the hypothesis that b is the result of random sampling from a population in which $\beta = 0$, the quantity

$$t = \frac{b}{\sqrt{\frac{s^2_{y.x}}{\Sigma(X-\bar{x})^2}}}$$

has the t-distribution with n-2 d.f. (Use 2-tailed t.)

10.12 Exact control of X. In the experiment on rice used as an example, it was possible to control X. In such cases, the arithmetic can be greatly reduced when it is desired to make test of significance only. An analysis of variance technique provides a sum of squares without making any apparent use of the X-values. A further extension gives information on the linearity of the relationship. Without explanation, the following analysis of variance is offered. The assumption that replicate differences are random and normally distributed is made. One would customarily remove these effects in an analysis of variance procedure.

Table 10.5

Source	d.f.	Sums of squares	Mean squares	F
Treatments	2	39,854.47		
Linearity	1	39,694.05	39,694.05	12.9
Deviations	1	160.42		
Residual	27	82,319.00	3,048.85	
Total	29	122,173.47		

The reduction due to linearity is calculated simply as

$$\frac{(3946 - 3055)^2}{2 \times 10} = 39,694.05$$

This is the reduction we obtained earlier with a great deal more effort. The advantage of being able to plan our experiment so as to reduce the arithmetic is obvious. In addition, information on the departure from linearity of regression (the X-values are equally spaced) is obtained. This can also be obtained by our regression procedure with still more computations.

Notice that deviations + residual add to the residual we worked with earlier. A test of significance is made without computing the regression line. This requires additional computations.

10.13 Inability to control X. With X subject to control, an experiment may be planned to cut down materially on the computations. In many experiments this is not possible. For example, yield and stand are two variates often measured; but rarely is stand controlled. In such cases, the arithmetic procedure outlined and exemplified in this chapter is generally necessary if the regression line is desired. An example appears later in this chapter.

In Chapters 12 and 13 regression is used in connection with experimental design and the analysis of the resulting data.

10.14 Linearity and the choice of X. In section 10.11, no use was made of the Y-observations for the mid-X in detecting linearity. In general, if a relationship is known to be linear, it is sufficient to observe Y-values for only two X-values. These X-values are best chosen as far apart as possible. This is obvious if you consider the distance between X-values and the variation of corresponding Y-means. Small sampling variation in $\bar{y}$'s can cause a considerable change in the sample regression coefficient if they are close together but the same size variation can have little effect when they are far apart. This is seen also from the sampling variance of b. Notice the $\Sigma(X-\bar{x})^2$ in the denominator. This quantity is equal to $\frac{(X_1-X_2)^2}{2}$ when only two X's are chosen. Clearly, s_b can be reduced by choosing the two X-values at a considerable distance from one another. Choosing only two X-values is putting all of one's eggs in a single basket and does not allow for detection of non-linearity. Thus it is common practice to use at least three X-values even where there is strong past evidence in favor of linearity.

10.15 Homogeneity of regression or the comparison of slopes. It often occurs that several estimates of a regression coefficient β are available and it is desired to test their homogeneity or to pool them into a single regression coefficient to give an estimate based on a larger number of observations. Thus, the rice data could supply 10 regression coefficients, one per replicate. We ignored replicates and obtained a single coefficient since a coefficient based on three pairs of observations would be of very little use. At other times, it might be desired to test for a common regression coefficient where two varieties were involved, or two seasons of a single variety. This subject will come up again in the analysis of covariance.

Recall that the regression coefficient is a linear combination of normally distributed variates. Consequently, a difference is also normally distributed. If we assume that the variates have a common variance, the t-test can be used to test whether the sample difference can be considered to have arisen by chance from a population of b's with mean zero. The test is given by

$$t = \frac{b_1 - b_2}{\sqrt{s_p^2 \left[\frac{1}{\Sigma(X_1-\bar{x}_1)^2} + \frac{1}{\Sigma(X_2-\bar{x}_2)^2}\right]}}$$

$$\text{where } s_p^2 = \frac{\left\{\Sigma(Y_1-\bar{y}_1)^2 - \frac{[\Sigma(X_1-\bar{x}_1)(Y_1-\bar{y}_1)]^2}{\Sigma(X_1-\bar{x}_1)^2}\right\}}{(n_1-2)+(n_2-2)}$$

$$+ \frac{\left\{\Sigma(Y_2-\bar{y}_2)^2 - \frac{[\Sigma(X_2-\bar{x}_2)(Y_2-\bar{y}_2)]^2}{\Sigma(X_2-\bar{x}_2)^2}\right\}}{(n_1-2)+(n_2-2)}$$

The subscripts refer to the sets of observations rather than the individual subscripts. Notice that the denominator is the pooled sum of squares for the Y-values after removing the part attributable to the independent variable. The divisor is the pooled d.f. The other factor is the sum of the reciprocals of the sums of squares for the fixed sets of X-values.

The analysis of variance procedure can be used. The appropriate test criterion is F. Table 10-6 indicates, by check marks, the required quantities. Line 1 consists of headings, line 2 involves the data associated with the first regression, line 3 that for the second. Line 4 is obtained by addition and gives the reduction in sum of squares when two regression lines are used. Line 5, columns 1, 2, and 3 are the addition of lines 2 and 3; columns 5 and 6 are found from the values of the earlier columns of this line. Here we have the reduction in sums of squares due to fitting a single regression to the data but correcting for two means. Finally, line 6, columns 5 or 6 (they give the same result) give the difference due to fitting two regression coefficients rather than one. The test of homogeneity is

$$F = \frac{\left[\frac{(\Sigma x_1 y_1)^2}{\Sigma x_1^2} + \frac{(\Sigma x_2 y_2)^2}{\Sigma x_2^2} - \frac{(\Sigma x_1 y_1 + \Sigma x_2 y_2)^2}{\Sigma x_1^2 + \Sigma x_2^2}\right]}{\left[\frac{\Sigma y_1^2 + \Sigma y_2^2 - \frac{(\Sigma x_1 y_1 + \Sigma x_2 y_2)^2}{\Sigma x_1^2 + \Sigma x_2^2}}{n_1 + n_2 - 4}\right]}$$

with 1 and $n_1 + n_2 - 4$ d.f. The square of t and the value F can be shown to be algebraically equivalent. Consequently, the tests are equivalent.

Table 10.6

Analysis of Variance

	1	2	3	4	5		6	7
1)	d.f.	$\Sigma(X-\bar{x})^2$	$\Sigma(X-\bar{x})(Y-\bar{y})$	$\Sigma(Y-\bar{y})^2$	$\frac{[\Sigma(X-\bar{x})(Y-\bar{y})]^2}{\Sigma(X-\bar{x})^2}$		Col.4)-Col.5)	d.f.
2)	n_1-1	✓	✓	✓	✓	(1 d.f.)	✓	n_1-2
3)	n_2-1	✓	✓	✓	✓	(1 d.f.)	✓	n_2-2
4)					✓	(2 d.f.)	✓	n_1+n_2-4
5)	n_1+n_2-2	✓	✓	✓	✓	(1 d.f.)	✓	n_1+n_2-3
6)					✓		✓	

-231-

The extension of the analysis of variance procedure to test the equality of more than two regression coefficients is obvious.

10.16 Bivariate distributions. In many cases, it is not possible to control X in the manner available to the rice experimenter. Often one must simply observe the X that exists and perform the extra computations described and illustrated in this chapter. Thus it is often not convenient to control the stand of a crop with more precision than given by a seeding rate. Or in measuring the heights and weights of adult American males, the individuals may be drawn at random and the pair of measurements observed. In an experiment to determine yield of potatoes as affected by fertilizer treatment, an observable X-value might be the rate of nematode infestation. These are but a few of many cases.

In some of these cases, there would seem to be no obvious choice for the dependent variate. In such cases, one is guided by the nature of the data and the use to which they are to be put. Consider the following pairs of observations on a single strain of guayule, a plant from which rubber is obtained. The variates are shrub weight and circumference of crown.

Table 10.7

Oven dry weight (grams)	Circumference of crown (cms.)
65	6.5
100	6.3
82	5.9
133	6.3
133	7.3
165	8.0
116	6.9
120	8.1
150	8.7
117	6.6

These paired observations are a portion from data which was obtained in a random manner. The individual plants were drawn at random from a field trial and a number of observations, including the two given here, made on each plant. There is nothing to distinguish any variable as an independent one.

When the paired observations are of this nature, i.e. are pairs of random variables, choose as dependent variable the one to be predicted. For these data, the sums of squares are 8120.9 $(\text{grams})^2$ for weight, 7.764 $(\text{cms})^2$ for circumference and the sum of cross products is 176.84 (gms x cms). To predict circumference, the reduction in sum

of squares and the reduced sum of squares are $\frac{(176.84)^2}{8120.9} = 3.851$ cms^2 and 3.913 cms^2 respectively with b = .022 cms./gm. To predict weight, the reduction in sum of squares and the reduced sum of squares are $\frac{(176.84)^2}{7.764} = 4027.9$ gms^2 and 4093.0 gms.^{2} respectively with b = 22.78 grams per cm. In this case, there are two regression equations. The two residual sums of squares are measured by perpendiculars to the different axes. Thus, different sums of squares are minimized in each case.

The case where there are two regression equations is easily distinguished from the case of one regression equation, regardless of which variable is to be predicted, by the manner in which the data have been collected and the use to which they are to be put.

The quantity

$$r = \frac{\Sigma(X-\bar{x})(Y-\bar{y})}{\sqrt{\Sigma(X-\bar{x})^2 \Sigma(Y-\bar{y})^2}}$$

called the correlation coefficient, is often calculated when such data are at hand. For these data,

$$r = \frac{176.84}{\sqrt{8120.9 \times 7.764}} = .70.$$

Notice that $100r^2$ is the percentage reduction in the sum of squares for either variable obtained by the use of the other as independent variable in connection with the moving average. This quantity is often called the coefficient of determination. The coefficient r will be discussed in the next chapter. The percentage reduction is a valuable quantity in that it is easily understood. With large samples, a small r may be statistically significant. However, if the sum of squares is reduced by only a small percentage, the value of the regression equation as a predictor may be subject to doubt. It is well if the experimenter can decide prior to the conduct of the experiment what sort of reduction will be meaningful. If this is done, one is not so often likely to permit his judgment to be carried away by algebra and arithmetic.

For example, if a new and cheap or quick method of determining the oil content of a product or vegetable is being compared with an old and expensive or time-consuming method, it is not sufficient to have a significant value of r; a high coefficient of determination

is necessary. In studies such as those concerning the many factors, genetic or environmental, which influence milk yield in dairy cattle, a small reduction of variation may be important. Here, a significant r may tell the story.

10.17 Prediction of X from Y. There are times when it is desired to predict a value of X from data in which the X's are not drawn at random. Workers in the field of dosage-mortality are continually doing this sort of thing. The only valid thing that can be done is to predict X by solving the prediction equation for X. Thus X is predicted by

$$X = \bar{x} + \frac{Y - \bar{y}}{b}$$

where for Y, we substitute that Y-value for which the X-value is required.

This is a point estimate of X and we usually require a confidence interval, i.e. a pair of estimates of X. The reader is referred to Eisenhart, C. : The interpretation of certain regression methods and their use in biological and industrial research, The Annals of Mathematical Statistics, vol. 10, 1939, pages 162-186.

10.18 Regression and design. The experimental designs and the corresponding simple calculations that have been discussed in chapters 8 and 9 are problems in regression where the arithmetic has been planned to be simply carried out. One case where the design enabled us to calculate the treatment and residual sums of squares without apparent recourse to the comparatively lengthy procedures of this chapter was given here.

You have, no doubt, noticed the comparison between the linear models of the analysis of variance and the linear model for regression. The one apparent difference is the lack of a set of observable parameters. These parameters are present but we do not bother to write them because they consist of either 0's or 1's.

Consider, for example, the case of observations on two treated plots and on two untreated plots. The analysis of variance involves a single d.f. for treatments and two degrees of freedom for error. The linear model is

$$X_{ij} = \mu + \rho_i + \epsilon_{ij}, \quad i = 1, 2; \; j = 1, 2.$$

where ρ is for treatment effect. This model may also be written in the form

$$X_i = \mu + \rho_1 W_i + \rho_2 Z_i + \epsilon_i, \quad i = 1, 2, 3, 4.$$

where 1 and 2 are associated with treatment 1, and 3 and 4 with treatment 2; and W_i takes on the value 1 for i = 1 and 2, and 0 for i = 3 and 4; and Z_i takes on the value 0 for i = 1 and 2, and 1 for i = 3 and 4. Clearly the models are equivalent.

The second model is seen to be similar to the one discussed in this chapter except that there are two independent variables instead of a single one. Also the subscript notation has been complicated in the second case which requires a lot of explanation. This is a problem in multiple regression, the subject of Chapter 13.

CHAPTER 11

LINEAR CORRELATION

11.1 Introduction. In Chapter 10, section 10.16, bivariate distributions were briefly discussed. Such distributions exhibit covariation. The name is descriptive since the joint variation of a pair of variates is involved. The sample measure of the covariation of two variables is the sum of the cross products of the deviations of the variables from their respective sample means, the numerator of r, a quantity called the correlation coefficient. The measure r is a dimensionless quantity, independent of the units of measurement of X and Y, unlike a variance or a regression coefficient. To calculate r, divide the sample covariance by the square root of the product of the sample variances, i.e. by the product of the sample standard deviations. Since multiplication of a variable by a constant leaves r unaffected, the correlation coefficient between height and weight remains constant regardless of whether height is measured in inches or centimeters or weight is measured in ounces or grams. Clearly, such a property is often desirable.

In the discussion of regression in the previous chapter, an average relation was involved which required that only one variable need be random and normally distributed. The regression coefficient, b, involved the units of measurement of the two variables and dealt primarily with means. Regression was again discussed in connection with bivariate distribution, but a new measure, covariance, was also introduced. For such data, variables that vary together perhaps because of external influences affecting both, covariance may seem to offer a more logical explanation than does a regression equation particularly when there is no obvious choice of dependent and independent variables. However, variances, which are dependent on the choice of a unit of measurement, affect it. The correlation coefficient is a measure of covariance for variables with unit variances, i.e. r is independent of the units of measurement; it is a measure of the intensity of association between two random variables. Here, then, is a third method for treating pairs of variables. The methods in the order in which they have been introduced are i) consideration of the variables separately, ignoring any relationship existing between them, ii) construction of a regression equation, and iii) examination of the correlation.

This chapter deals with correlation but we must not lose sight of the fact that an adequate explanation of data from bivariate distributions must involve two means, two variances, and the correlation coefficient.

11.2 Definitions and explanation. Correlation is a measure of the degree to which variables vary together and is defined as follows:

$$r = \frac{\Sigma(X-\bar{x})(Y-\bar{y})}{\sqrt{\Sigma(X-\bar{x})^2}\sqrt{\Sigma(Y-\bar{y})^2}}$$

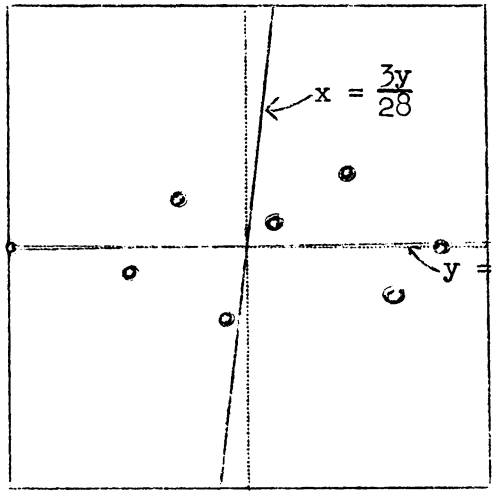
Thus, r is a measure of the association of jointly varying variables. An examination of Figure 11.1 will show what is involved. The data for these figures are manufactured to have desirable properties including $\bar{x} = 0$, $\bar{y} = 0$ so that regression lines pass through the origin. In figure a), the individual points seem to cluster about the X-axis. This is because the variance of X is larger than that of Y. The clustering does not suggest covariance. For these data, $s_x = 6$, $s_y = 2$, $\Sigma xy = 3/7$. The regression of Y on X is given by $Y = (1/84)X$. The regression of X on Y is given by $Y = (3/28)X$. Recall that a different sum of squares has been minimized for each equation. (These equations are not intended to be meaningful except for the interpretation of r.) In each case, the regression line is close to an axis. For figure a' the data of a) have been divided by appropriate standard deviations so that the variables have unit variance. Now the points show no tendency to cluster about any line, axis or other. This lack of cluster when the points have a common variance, unity here, is typical of data showing little or no correlation. Regression lines for the new set of data are $Y = (1/28)X$ and $X = (1/28)Y$. The regression coefficients are now the same, namely the value of the correlation coefficient.

Actually, these data show a negligible positive covariation and the correlation is $r = 1/28 = .036$. Data with unit variance for each variable, showing small negative correlations would yield regression equations close to the same axes respectively, but in the other pair of quadrants. Notice that r measures an angle where the standard deviations are the units of measurement just as b measured an angle dependent on the units of measurement of the original data. For large amounts of data, the pairs of observations plot in a circular

X	Y
-10	0
-5	-1
-3	+2
-1	-3
+1	+1
+4	+3
+6	-2
+8	0
$\bar{x}$: 0	0
s^2 : 36	4
s : 6	2

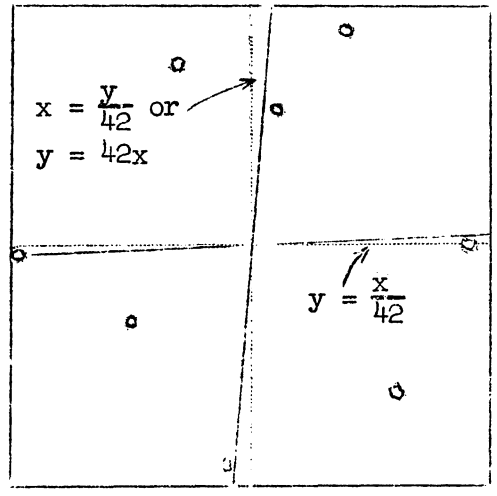
CoV: 3/7

a)



r = 1/28 = .036
y = [3/(7 x 36)]x = (1/84)x
x = [3/(7 x 4)]y = (3/28)y

a'



X	Y
-1 2/3	0
- 5/6	- 1/2
- 1/2	+1
- 1/6	-1 1/2
+ 1/6	+1
+ 2/3	+1 1/2
+1	-1
+1 1/3	0
$\bar{x}$: 0	0
s^2 : 1	1
s : 1	1

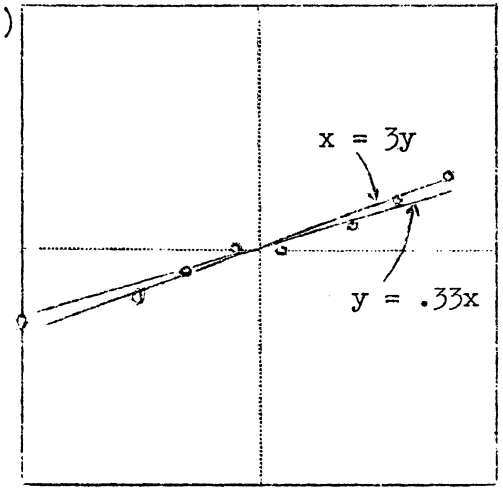
CoV = 1/42

r = 1/28 = .036
y = (1/28) x
x = (1/28) y

X	Y
-10	-3
-5	-2
- 3	-1
- 1	0
+ 1	0
+ 4	+1
+ 6	+2
+ 8	+3
$\bar{x}$: 0	0
s^2: 36	4
s: 6	2

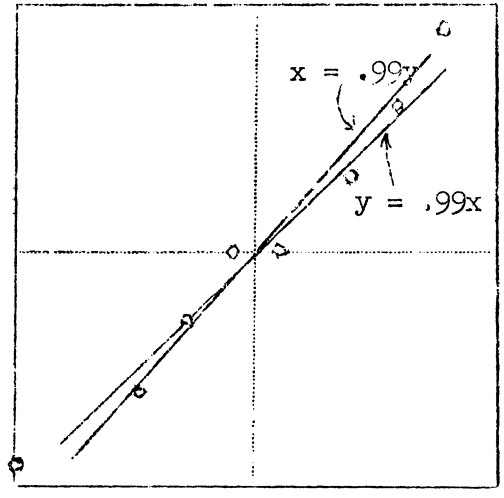
CoV = 83/7

b)



r = .99
y = [83/(7 x 36)] x = .33x
x = [83/(7 x 4)] x = 3.0y

b'



X	Y
-1 2/3	-1 1/2
- 5/6	-1
- 1/2	- 1/2
- 1/6	0
+ 1/6	0
+ 2/3	+ 1/2
+1	+1
+1 1/3	+1 1/2
$\bar{x}$: 0	0
s^2: 1	1
s: 1	1

CoV = 83/(7 x 12)

r = 83/(7 x 12) = .99
y = .99x
x = .99y

-3-
Figure 11.1

or elliptical form, the ellipse having axes roughly parallel to the plotting axes.

Figure b) makes use of the same numbers as a) but the X- and Y-values are now paired in a different manner. Here $\Sigma xy = 83/7$ while the means and variances necessarily remain the same. The regression equations are $Y = .33X$ and $X = 3Y$ and the lines are nearly coincident, lying closer to the X-axis than to the Y-axis. In b' where the data of b) have been given unit variance, the regression equations are $Y = .99X$ and $X = .99Y$. The regression coefficients are now the same and the pair of regression lines are both close to the 45° line. Here we have pronounced covariation and a high positive correlation, $r = .99$. High negative correlation would give a pair of lines near the 45° line through the other pair of quadrants.

Large amounts of data which show high correlation, either positive or negative, plot in the form of an ellipse. If these data have variates with unit variances, then the axes of the ellipse will be roughly at an angle of 45° to the plotting axes. One is generally fortunate enough to see a similar tendency, except possibly for the 45° angle, in the raw data but an unfortunate choice of unit on the plotting paper will tend to obscure the relationship. Consequently, one must look critically at the variances while examining such a chart.

It can be shown that r lies between plus one and minus one, i.e. $-1 \leq r \leq +1$. The values of plus one and minus one indicate a perfect correlation or an exact mathematical relationship between the pair of variables. This takes them outside the province of statistics. Such quantities are encountered by the experimentalist rarely other than in moments of carelessness. Thus, one would obtain a perfect correlation, except for rounding errors, if one correlated height of an individual with shoulder height plus shoulder to top of head; or if a straight line were fitted to a pair of points. The fact that rounding errors could lead to a correlation coefficient near one in absolute value, for such data, makes extremely high correlations suspect. Correlations greater in absolute value than one are generally due to computational or rounding errors. They aren't real.

Note that

$$r^2 = \frac{(\Sigma xy)^2}{(\Sigma x^2)(\Sigma y^2)} = \frac{\Sigma xy}{\Sigma x^2} \cdot \frac{\Sigma xy}{\Sigma y^2} = (b_{y.x})(b_{x.y})$$

where $b_{y.x}$ and $b_{x.y}$ are the coefficients for the regressions of y on x and x on y respectively.

Two quantities closely related to r which you may encounter in your reading are the coefficient of determination and the coefficient of alienation. The coefficient of determination is r^2, the square of the correlation coefficient. It is useful in regression problems rather than correlation problems since it gives a measure of the reduction in the sum of squares of the dependent variable, due to the independent variable. I.e., it is the proportion of the sum of squares for which there is an explanation in a covariate. It is very nearly the proportion of the variance that has been explained but because of the difference in the d.f. between the total variance and the variance about regression, a difference of one, it is not exactly so. It may be calculated directly from the analysis of variance table for regression as

$$r^2 = \frac{\left[\frac{(\Sigma xy)^2}{\Sigma x^2}\right]}{\Sigma y^2} = \frac{(\Sigma xy)^2}{(\Sigma x^2)(\Sigma y^2)},$$

i.e. the ratio of the reduction in the sum of squares to the unreduced sum of squares for the dependent variable. This ratio of two sums of squares is a multiple of the ratio of two variances, the multiple being the ratio of the d.f.

The quantity $(1-r^2) = k^2$ is the portion of the sum of squares of the dependent variable that has not been explained. This is sometimes called the coefficient of non-determination. Its square root, k, is called the coefficient of alienation. The quantity $k' = 1 - k$ has been called the improvement factor. It is doubtful if you will use any of these terms but they are presented for completeness.

11.3 Correlation and regression. The distinction between correlation and regression has already been pointed out. Correlation measures a co-relation, a joint property of the variables. Neither variable is claimed as dependent or independent. The pair of observations is drawn at random from a bivariate distribution. Regression deals with a relationship between the mean of a random variable

and an independent variable. Here there is a dependent and an independent variable when the data are to be interpreted though the collection of the data may not have been carried out as that implies. That is, randomness in the independent variable is not required but may be present. The correlation coefficient is not affected by the unit of measurement whereas the regression coefficient is.

The quantity r can be useful in regression problems in that it measures the fraction by which a sum of squares is reduced. Thus,

$$r^2\Sigma y^2 = \frac{(\Sigma xy)^2}{(\Sigma x^2)(\Sigma y^2)} \cdot \Sigma y^2 = \frac{(\Sigma xy)^2}{\Sigma x^2}$$

the reduction in the sum of squares, i.e. the part of Σy^2 attributed to the independent variable. Where two regression coefficients are calculated, the following arithmetical relationship has been observed:

$$r = \sqrt{b_{y.x} b_{x.y}}$$

11.4 Confidence statements and tests of hypothesis. Like any other sample value, r is a variable with a sampling distribution. This distribution is symmetric about zero when the population value of the correlation coefficient, denoted by ρ, is zero. For large samples, the distribution approaches normality.

When $\rho \neq 0$, the distribution of r is markedly skew and approaches normality much more slowly than when $\rho = 0$. This, of course, depends on how different from zero, ρ is. A normal approximation may not be very accurate.

To get around the difficulties of anormality, a special table or some transformation of r which makes normal theory applicable is required.

The simplest method for the setting of confidence limits is supplied by Pearson and David's Tables of the ordinates and probability integral of the distribution of the correlation coefficient in small samples. Easy-to-use charts which are sufficiently accurate for most of us are available.

An alternative which is sufficiently accurate and convenient is to compute the variable $z = .5 \ln(1+r)/(1-r)$ which is approximately normally distributed with approximate mean and standard deviation of $.5 \ln(1+\rho)/(1-\rho)$ and $1/\sqrt{n-3}$ respectively regardless of the value of

ρ. (ln refers to the natural logarithm, or log to the base e.) Use z-values from the normal tables or the bottom line of a t-table. Confidence limits are set for z and converted to ρ. Thus for our artificial example with r = .99, we have n-3 = 5 and $s_z = \sqrt{1/5} = .45$, $\ln(1+r)/(1-r) = \ln(1.99/.01) = \ln 199 = 5.293$ and .5 ln 199 = 2.65 = z. Now 95% confidence limits are given by $2.65 \pm (1.96)(.45) = 1.77$ and 3.53. The confidence limits are .94 and 1.0. To facilitate the computation of z from r and r from the limits on z, see Snedecor, Figure 7.4.

To test the hypothesis that ρ, the population correlation coefficient, equals some specified value, the normal distribution is used as implied in the previous paragraph. If the hypothesis is that $\rho = 0$, then a simpler procedure may be used. For $\rho = 0$, we use

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$$

This square of t can be shown to be algebraically equivalent to F as calculated from an analysis of variance for regression and, consequently, to t for testing $\beta = 0$ in a regression problem. This test cannot be used for testing $\rho = k \neq 0$, i.e. some constant other than zero. In many practical problems, a test of an hypothesis about β will be the appropriate one. Once again, make a distinction between significance and meaningfulness. This is not a serious problem with small samples but can be one with large samples. In large samples, small values of r may be significant. However, if the percentage reduction in variance is small, the correlation may be quite useless. The test of a difference between two values of ρ was implied several paragraphs previously. The two r's are converted to z's and the usual normal test of section 5.7 applied. It is to be noted that as far as the tests of r are concerned in themselves, z-values must be found but no reconversion is required as in the case of confidence limits.

In Chapter 10, it was shown that a test of homogeneity of b-values, i.e. of the hypothesis of a common β, involved the difference between the reductions in sums of squares of the independent variable when a single coefficient was used and when several were used. If the result indicated homogeneity then the single coefficient was used and its calculation was obvious.

In some cases, it is desired to test the homogeneity of correlation coefficients and to obtain a single coefficient if there seems to be homogeneity. For example, measurements on two characteristics of a crop or of some type of animal might be available for several strains or breeds. It is quite likely that the variances involved from strain to strain or breed to breed will not be homogeneous so that a simple pooling of the data and calculation of a single correlation coefficient is not valid. Or, it is possible that one is simply interested in a pooled value of r as a measure of co-relation and has the correlation coefficients available. The test of homogeneity and the method of pooling are illustrated in Table 11.1. A description of the process follows.

Table 11.1

Strain	No.	r_i = correlation coefficient	$z_i = (1/2) \log \frac{1+r}{1-r}$	$z_i - \bar{z}_w$	$(z_i - \bar{z}_w)^2 (n_i - 3)$
405	50	.362	.379	-.0913	.392
407	50	.419	.446	+.0243	.028
416	50	.527	.586	+.1157	.629
	150				$\chi^2 = 1.049$

$$\bar{z}_w = \frac{\Sigma(n_i - 3)z_i}{\Sigma(n_i - 3)} = .4703$$

The test criterion is χ^2 with one d.f. The value is clearly not significant here. Correlations are for percent resin content and percent rubber content in guayule.

The z_i are normally distributed with mean and variance as previously stated. A weighted mean, $\bar{z}_w$, is required, the weights being the reciprocals of the variances, namely $n_i - 3$. In the resulting sum, those r's based on larger numbers of observations have greater importance. The divisor is the sum of the weights. Note that the individual $(z_i - \bar{z})$'s must be calculated.

There is a small bias in z. As a result, this may be serious if large numbers of correlations are averaged. Since only three are involved here, we convert $\bar{z}_w = .4703$ to r = .438 without hesitation. For the method of eliminating the bias, see Snedecor, p. 154-5.

11.5 Common elements and other matters. A convenient device offering some help in the understanding of correlation is found in common elements. It is useful when both observations of the pairs have the same unit of measurement. It is not an explanation of correlation and has considerable limitations in interpretations.

To construct an example with common elements, proceed as follows. Make up a table with two columns, one for X and one for Y. Obtain two or three numbers at random from the normally distributed data of Chapter 4 and place them in both columns. Now draw a pair at random and put one in each column. The two column sums constitute a pair of observations. Proceed with the sampling as described, obtaining a sequence of pairs of observations. You have now obtained a sample from a population in which the correlation coefficient is

$$\frac{n_{12}}{\sqrt{n_1 n_2}}$$

where n_{12} is the number of common elements, two or three were suggested, and n_1 and n_2 are the number of elements in each sum, a single additional number was suggested for both X and Y so that n_1 and n_2 are equal for our case. Now, if the sample is not too small, a sample correlation coefficient of approximately the same magnitude as that of the population should be obtained. In this case, it is clear that the population correlation coefficient measures the fraction of common elements.

You can probably think of examples in your field where you are prepared to consider common elements as offering a rough explanation for a correlation. Thus, a brother-sister correlation in humans or animals for some multi-factorial trait where the factors are assumed additive, might be explained in a crude sort of way by common elements. Obvious difficulties are the equal additions and additivity. And there's still the problem of interpreting the common element.

Common elements imply nothing about cause and effect relationships. This is as it should be. They do, however, serve to point out that correlation and regression are different matters.

Regression can be related to the common elements approach by drawing all X_1 elements and then transforming some of them, drawn at random, to X_2, finally adding extra ones to the transferred data to

obtain X_2. Now we can say that values of X_2 are controlled, in some degree, by the values of X_1. The emphasis has been shifted from correlation to regression, though clearly r will have the same value no matter which approach we use to obtain the data. Now we may calculate a regression equation, namely,

$$\hat{X}_2 = \mu_2 + \frac{n_{12}}{n_1} (X_1 - \mu_1),$$

where $\frac{n_{12}}{n_1}$, like μ_1 and μ_2, is a population parameter. Notice that the regression coefficient is the ratio of the number of elements transferred from X_1 to X_2 to the number in X_1. The number of elements in X_2 does not enter into our equation. Also, it is apparent that a single regression equation is valid.

Studies involving correlation, like other studies, should be undertaken on an intelligent basis. No doubt there exists a correlation between the yield of wheat in North America and the infant mortality rate in some backward country. After all, North America is a growing country and has continually increasing requirements for wheat while many backward countries are now receiving medical attention that was unheard of a decade or two ago. Even if the correlation is non-linear, there is a good chance that some country can be found where there are a few rates extreme enough that the calculation of r will lead to a high value. Such correlations as these are often called spurious or nonsense correlations. This term is used, particularly when time trends exist, by economists. Unfortunately, you will find that just as meaningless examples have crept into the literature of various sciences. More specifically, spurious correlations occur when the correlation coefficient between two variables is due, to some extent at least, to factors external to those which produce the supposed correlation. In the above example, an external factor could be time.

For some experimentalists, this is simply the problem of heterogeneous data. Snedecor considers the three variables, cob circumference, number of kernel rows, and ear circumference. The correlation between cob circumference and number of kernel rows is .507. However, if one considers only ears having the same circumference, then the correlation between these variables is only .105.

We shall consider this type of problem again in a chapter on multiple and partial correlations. Meanwhile, you will notice that spurious correlations need not be worthless correlations.

11.6 Intraclass correlation. It sometimes happens that a correlation coefficient is desired where the members of the pairs of variables are not easily designated as X_1 and as X_2. For example, we might have the production records of twin cattle. Which of the records shall we denote by X_1? Clearly, we can use a random procedure but then we have a sampling variation over and above that due to the selection of the twin animals. One gets around this difficulty by considering the two possible pairings, i.e. use cow A for the X_1 value and cow B for the X_2 value, then cow B for the X_1 value and cow A for the X_2 value. In another case, we might choose 3 males from each of many litters of mice. In this case, one obtains $3 \times 2 = 6$ entries in the correlation table for each pair of values. And there will be cases where we have unequal numbers of animals from group to group.

One does not customarily do the work implied by the previous paragraph but obtains a value of the coefficient from the calculation of certain variances. The formula, given below, can be shown to yield the same answer as that implied above. The formula is

$$r = \frac{\text{(M.Sq.(classification) - M.Sq.(error)}}{(k-1)\text{M.Sq.(error)} + \text{M.Sq.(classification)}}$$

These mean squares are obtained from an analysis of variance for a one-way classification, the classification being the k families in our animal experiments.

Notice the relationship with components of variance. Referring to Chapter 8 for the one-way classification, observe that from the mean square column in an analysis of variance, estimates of σ_m^2 and σ^2 are available. The ratio

$$\frac{\hat{\sigma}_m^2}{\hat{\sigma}_m^2 + \hat{\sigma}^2}$$

where $\hat{}$ denotes estimate, gives the numerical answer and the formulas can be shown to be algebraically equivalent. This is a ratio of two variances and may not look much like a correlation to you. However, if you will look at the last equation of section 11.2, you will see

that a correlation coefficient has already been shown to be a multiple of a ratio of variances.

While the value of an intraclass correlation may go as high as +1 in theory, when $\hat{\sigma}^2$ is zero, it cannot drop below -1/(k-1). For this reason, we modify the word correlation. The ordinary correlation is sometimes called the product-moment correlation to emphasize its nature; it is also called the interclass correlation.

A test of significance for an intraclass correlation is now obvious. We simply test for the presence of the component σ_m^2. The criterion is F and the numerator and the denominator are the values from the analysis of variance table. Consequently, one usually performs a test of significance before estimating r.

11.7 Paired observations. In Chapter 5, the problem of testing pairs of means was discussed. The discussion included tests when the numbers of observations in each mean were equal and paired. The case of the paired observations was again discussed in Chapter 9 on the analysis of variance. Correlation supplies us with another technique for testing the difference between a pair of means, i.e. of the hypothesis that the populations have a common mean, and points out a relationship between analysis of variance and correlation as well as the fact that a well-designed experiment can save the experimentalist considerable arithmetic.

Consider a set of pairs of observations. The pairing implies that the members of the pair are, in some way, more closely related than a pair whose members are drawn at random, one from the X_{1i} values and another from the X_{2i} values; in other words, the existence of a correlation between the members of the pairs is implied. Let us calculate this correlation for the first example of section 5.9. $\Sigma xy = 47.29$ and $r = .680$. The null hypothesis is that $\mu_1 = \mu_2$. If it can be said that $\bar{x}_1 - \bar{x}_2$ is greater than zero by an amount that cannot be explained by random sampling from a normal distribution of differences with mean zero, then we conclude that the means are not a random sample from that population and, consequently, that $\mu_1 \neq \mu_2$. The appropriate criterion is t, since we do not know the population value of the variance. Theory tells us that the variance of a difference is given by

$$\sigma_1^2 + \sigma_2^2 - 2\sigma_{12} = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2.$$

We have estimates of σ_1, σ_2, and of σ_{12}, the covariance. Define $\rho = \frac{\sigma_{12}}{\sigma_1 \times \sigma_2}$ and, as a result, $\sigma_{12} = \rho\sigma_1\sigma_2$. Substituting in the above equation, we find $s^2_{x_1-x_2} = \frac{71.66 + 67.42 - 2 \times 47.29}{9} = 4.944$ and $s^2_{\bar{x}_1-\bar{x}_2} = .4944$ as in Chapter 5.

11.8 Non-linear correlation. As this is to be the subject of a later chapter, we shall do little more than indicate that it exists. In this chapter, we have dealt with linear correlation only. No doubt, you will have pairs of observations, which when plotted appear to fall on a curved line rather than on a straight line. You may even have an hypothesis that this is to be expected. Now you will want to use some curvilinear equation to express the relation between the members of your paired observations. You may still want a measure of how closely the points cluster about this line. A measure is supplied in the multiple correlation coefficient. Its calculation is not generally as simple as that of the linear correlation coefficient. At times, a curvilinear relationship is made linear by changing the scale of one of the variables. Thus, if a theory required that a relationship be of the form

$$Y = aX^b,$$

then a linear relationship can be obtained by taking logarithms to give

$$\log Y = \log a + b \log X$$

It may be desirable to make such a transformation for computational purposes. Sometimes this will not be appropriate for making tests of hypotheses or for setting confidence limits on parameters. These matters will be further discussed in a later chapter.

CHAPTER 15

ENUMERATION DATA

15.1 Introduction. Most of the previous chapters have dealt with measurement data with such units as inches, square centimeters, bushels, pounds, etc., measures of length, area, volume, weight, and so on. Obtaining these data required the use of some standard unit of measure and a measuring device. Enumeration data arise when dealing with the presence or absence of attributes. Such data consist of the numbers of individuals falling into stated classes. Thus, the geneticist samples a population of mice and observes the number of males and females or he samples a population of individuals to learn how many can roll their tongues and how many can't. The pollster classifies the people in his sample according to the way they claim to be going to vote and notes the numbers in each class. A production man may count the number of defective and non-defective units in his production. A person running a taste-testing experiment considers the number of people who claim or do not claim to distinguish between pairs of samples of some food product.

Two is not the only possible number of categories for enumeration data and, clearly, measurement data may be collected as a matter of convenience, though with a loss of information, as enumeration data. For example, a sample of people may be requested to supply information as to whether their income is less than $2,500, between $2,500 and $7,500, and over $7,500.

15.2 The Problem. The problem is the usual one, the drawing of valid conclusions. One collects a sample and then wishes to make inferences about a population. If an inference is to be valid, some thought and care must be given to the collection of the sample. The sample must be drawn from the population about which the inference is to be made and must be representative of that population. This is usually achieved by drawing a random sample from the required population. Sampling variation is taken into consideration when making the inference about the sampled population. The sampler must also be prepared to accept the fact that he will make some incorrect inferences since he must expect to get some unusual samples.

An inference may take the form of a point estimate of a parameter, a confidence interval for a parameter, or the acceptance or rejection of some hypothesis. The problem is essentially that of any of the previous chapters. All that is required at this stage is appropriate estimates and tests.

15.3 Estimation of a percentage and a ratio. There are many situations where data naturally falling into only two categories are collected solely for the purpose of estimating a percentage. Opinion polls and some farm survey data are only two of many situations where this may be so.

English et al. sampled, among other people, 302 persons without coronary disease and classified them as smokers and non-smokers, 187 and 115 respectively. Suppose it is required to set confidence limits on the percentage of non-smokers in the population of persons without coronary disease. Clearly, there is a population value for the probability of selecting a non-smoker. This does not say that there is a probability that a person without coronary disease is a non-smoker. A person either is or is not a non-smoker.

The sample of persons without coronary disease yields an estimate of the required probability. It is required to place limits about the estimate with some stated degree of assurance that the interval will contain the parameter. This requires the distribution of the estimate, which in turn requires some knowledge about the distribution of the individuals.

If sampling is random and the population infinite, then the problem is one involving the binomial distribution, i.e. where every observation falls in one of two possible categories. This was discussed in Chapter 3. It is customary in most situations such as this to place upper and lower limits on an estimate in such a way that our incorrect statements, i.e. sets of confidence limits that do not contain the parameter although it is claimed that they do, will err equally in both directions, i.e. that the parameter will be above the limits as often as it is below them in the incorrect statements. The two values of p that constitute the limits are an upper one such that if it were the true value then the pair of numbers observed, viz. 115 and 187, would be the last pair acceptable as being due to random sampling from a population with this parameter, and a lower one such that if it were the true value then the pair observed would be the last pair acceptable at the specified probability level. In the first case, 114 and 188 would not be acceptable and in the second case, 116 and 186.

Tables of the Binomial Probability Distribution (National Bureau of Standards, U.S.A., Applied Mathematics Series 6), have been prepared for sample size values of n up to 50 and values of p, the probability of the less frequent event, from .01 to .50 by steps of .01. These tables include probabilities of individual terms and sums of probabilities of consecutive terms from 0 to 49 by steps of 1. Other tables are also available. Such tables are useful but more convenient tables are also available. Snedecor's Table 1.1 contains both 95% and 99% confidence limits, given as percentages, for a limited number of sample sizes. If you look opposite 115/302 = .38, under sample sizes 250 and 1000, you will find 95% confidence limits given as 32-44 and 35-41 respectively. Clearly we won't be much in error if we accept 32-44 and the error will be the conservative one.

If you use Clopper and Pearson's Confidence Belts for p, Dixon and Massey, Tables 9, you should be able to do about as well. Your authors obtained 32-45 by reading the n = 250 line and not trying to interpolate. To use this figure, obtain the ratio x/n, .38 here, draw a line vertical to the lower axis at this point, then draw lines parallel to the lower axis from the intersections of the vertical line with the two lines marked 250. From the column scale, read the pair of limits. Be careful that you identify the pair of 250 lines correctly.

Finally, you may calculate the confidence limits. You will probably decide to claim that your estimate of p is normally distributed with variance $p(1-p)/n$ and will use your estimate of p in calculating this variance since you are not hypothesizing a value for p. The 95% confidence limits for the percentage are given by

$$.381 \pm 1.96\sqrt{\frac{.38 \times .62 \times 100^2}{302}}$$

or .33 to .44. This is not the only possible approximation using the assumption of normality but it is convenient and probably differs little from any other approximation.

Table 15.1 gives sample sizes for values of a number of observed proportions for which normal theory applies.

Table 15.1

Observed Proportion	Sample size for normal theory to apply
.4 - .6	50
.3 - .7	100
.2 - .8	400
.1 - .9	1000

By now, you should be fairly well convinced that any of the above approximations is adequate. For sample size of less than 50, Snedecor's Table 1.1 and the Clopper-Pearson charts of Dixon and Massey are very adequate. For samples of size greater than 50, binomial probability paper (not shown here) looks pretty good if you have many to obtain but you may be prepared to go through the computations as illustrated if you have only a few. You will want to remember that the binomial distribution is discrete and not continuous. If you have very large numbers in your samples, your binomial model will probably not be an exact one but an approximate though very useful one. Thus, use of the normal approximation will be no more serious than the assumption of a binomial distribution.

15.4 Test of goodness-of-fit. In many cases, especially where genetic ratios are involved, one has an hypothesis about what an appropriate ratio is. In such cases, one wishes to test whether or not the random sample can be considered to be from the specified population. If confidence limits are calculated and contain the hypothesized ratio, then there is no significance at the stated confidence level and vice-versa.

In a certain F_1 generation of Drosophila melanogaster, 35 males and 46 females were obtained. It is required to test whether the departure from a theoretical 1:1 ratio can be attributed to chance alone.

Probably the most common test criterion for such problems is Chi-square. The chi-square criterion is defined by the equation

$$\chi^2 = \sum_{i=1}^{n} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} .$$

Expected refers to the average number to be expected when the hypothesis is true, i.e. the number called for by the theory; n is the number of cells in which observations appear. In our case, n = 2.

In theory, chi-square is defined as the sum of squares of independent and normally distributed variables. In our case, there are two such quantities to be squared and summed but you will find them numerically equal though opposite in sign. This is a general characteristic and not peculiar to our data. Consequently, we can claim only one d.f. for our chi-square. This can also be explained

in the following manner. We can assign any number we wish to one of the two cells but the second number is determined by the fixed total, the sample size. Alternately, you may wish to say that you have placed a restriction on your deviations, namely that their sum be zero. In either case, one of the d.f. is used.

For the Drosophila data,

$$\chi^2 = \frac{(35 - 40.5)^2}{40.5} + \frac{(46 - 40.5)^2}{40.5} = 1.49$$

Tables of chi-square for one degree of freedom show that the probability of obtaining a χ^2 value as high or higher than this, when the ratio is 1:1, is between .30 and .20. Clearly, deviations from the 1:1 ratio can be explained by chance.

The chi-square test is an approximation. You will have noticed this since it was said that chi-square was composed of a sum of squares of independent and normally distributed variables while our data should follow a binomial rather than a normal distribution. The normal deviate which we are using has mean .5 (this was our hypothesis) and variance $pq/n = (.5)(.5)/81$. If you wish to use normal tables directly instead of a chi-square table, calculate

$$z = \frac{(35/81 - .5)}{\sqrt{(.5)(.5)/81}}$$

$$= 1.22.$$

Its square is 1.49 as we have already seen.

Chi-square lends itself conveniently to pooling the results of a number of tests. This will be illustrated later in this chapter. It is also convenient arithmetically in that no square root is required.

A word of review about tests of significance. When data have been obtained and an hypothesis tested, the value of the criterion may not depart significantly from the values to be expected when the null hypothesis is true. Consequently we accept this hypothesis. This is not to say that we have evidence in favor of the hypothesis but rather that we have no significant evidence against it. In other words, there is not sufficient evidence to lead us to formulate another hypothesis. This is in keeping with a desire to avoid a possibly complicated hypothesis, even though it may explain the data satisfactorily, so long as a simple hypothesis will do so.

15.5 Tests of several hypotheses. The case of testing an hypothesis about a ratio where the alternative hypothesis was only that the ratio was other than that specified by the hypothesis has just been discussed. In many genetic problems, the experimenter will have several ratios in mind with no reason to prefer one rather than the other. For example, he may be in doubt as to whether the theoretical ratio should be 3:1 or 1:1 but will have no reason to favor one rather than the other. Consequently, he will have no reason to consider an error of one type as more to be avoided than an error of the other; since there is no basis for saying which error should be referred to as Type I; to claim the population ratio is 3:1 when it is 1:1 is as serious as to claim it is 1:1 when it is 3:1. The use of chi-square with a fixed probability of Type I error can lead to serious Type II errors for small samples. When more than two hypotheses are to be chosen among, chi-square is not satisfactory. A partial solution of this problem is given here for two or three ratios. It ignores the fact that linkage interferes with the independent assortment of genes, an assumption used in developing the test regions given here.

Suppose an individual falls into one of two distinct classes about which we have only two hypotheses, namely that their theoretical ratio is 1:1 or 3:1. A common test procedure is to select one of the ratios, say 3:1, as the null hypothesis and to fix the probability of claiming the ratio to be 1:1 when it really is 3:1, at some value such as 5% or 1%. Although the test criterion is usually chosen to minimize type II error, the procedure tells us nothing obvious about its magnitude. For example, with a sample size of 20, if we accept the ratio of 1:1 when there are not more than 11 in the group associated with 3 of the 3:1 ratio and accept the 3:1 ratio when the number goes over 11, then we make the wrong decision if 3:1 is the true hypothesis with probability .0409 but we make the wrong decision if 1:1 is the true hypothesis with probability .2517. This regards 3:1 as the null hypothesis. On the other hand, if the acceptance and rejection regions are changed so that we accept the 1:1 ratio when the number falling in the class associated with the 3 of the 3:1 ratio is not more than 13, and reject the 3:1 ratio when there are more than 13 in this class, we make the wrong decision when

the true ratio is 1:1 with probability .0577 and the wrong decision when the true ratio is 3:1 with probability .2142. This regards 1:1 as the null hypothesis. Now, if we decide in favor of the 1:1 ratio for up to and including 12 in the first class, we make the wrong decision when 1:1 is the true ratio with probability .1316 and the wrong decision when 3:1 is the true ratio with probability .1018. This would seem to be the most equitable procedure if we have no reason to be more afraid of one type of error than the other. The results are summarized in Table 15.2.

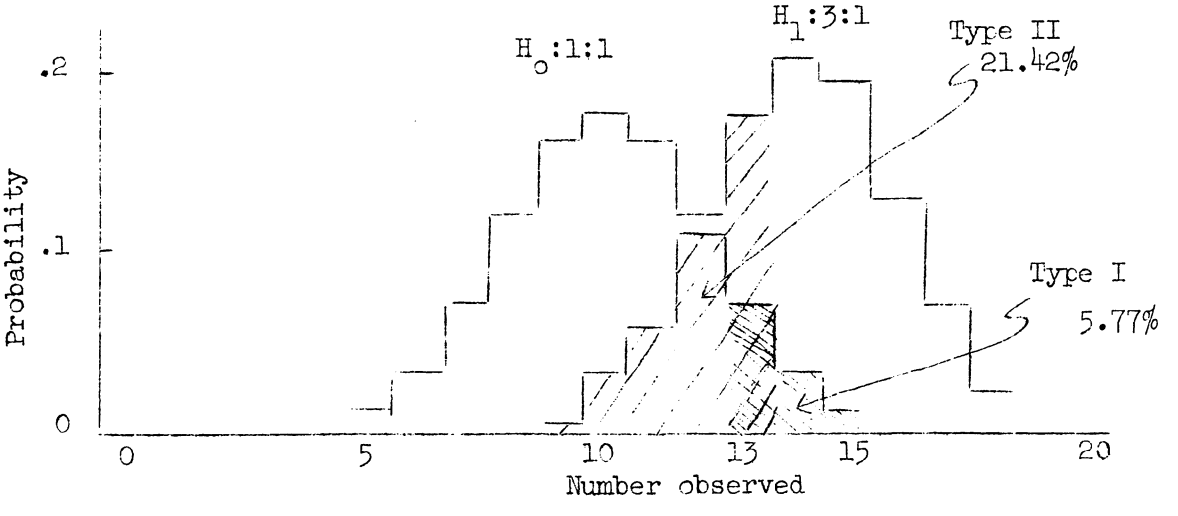
Table 15.2

Probabilities of making wrong decisions for ratios 1:1 and 3:1, n = 20

Acceptance regions for ratios:		Probabilities of making a wrong decision when the true ratio is:	
1:1	3:1	1:1	3:1
(Number in first class)			
0-11	12-20	.2517	.0409
0-12	13-20	.1316	.1018
0-13	14-20	.0577	.2142

Figure 15.1 may be of some help in appreciating what is involved. Regardless of what the true ratio is, any sample from (0,20) right through (20,0) is possible. The various samples have different probabilities associated with them according to the true ratio. The two probability distributions are shown in the figure in histogram form for visual convenience. The probabilities are more correctly shown as vertical bars without width at each of the integral values 0, 1, ..., 20. In Figure 15.1, the decision is made in favor of the ratio 1:1 for not more than 13 in the first class and for the ratio 3:1 otherwise.

Figure 15.1



The method just discussed states that it is not always desirable to fix the Type I error in advance. This is particularly so for genetic ratios where there is no reason to choose one of the ratios rather than the other, to be associated with the null hypothesis. The method does suggest that one minimize the maximum probability of an error of any kind. This gives us what is called a minimax solution. In Table 15.2, the maxima are .2517, .1316, in column 1, and then .2142 in column 2. The minimum of these maxima is .1316 which leads us to choose 12 as the last number to be associated with the decision 1:1 and 13 as the first to be associated with the decision 3:1. We will be wrong, at most, 13.16% of the time with this rule, and wrong this often only if we are never presented with data except from a population where the ratio is 1:1.

Table 15.3 contains minimax solutions for various sample sizes chosen so as to include values giving near the usual 5% and 1% probabilities. The probability of an error in decision when the other ratio is the true one, is also given.

Table 15.3

Sample sizes and Probabilities of making wrong decisions for ratios 1:1 and 3:1

Sample size	Accept ratio	Acceptance regions	Probabilities of wrong decisions
44	1:1	0-27	.0481 (cf. .05)
	3:1	28-44	.0316
80	1:1	0-50	.0089 (cf. .01)
	3:1	51-80	.0000
Dividing line (regardless of sample size) R /n ≐ .631			

Probabilities of exactly .05 and .01 cannot be obtained since we are dealing with a discrete distribution.

Chi-square is useful in deciding between two hypotheses in some cases. However, it is inappropriate when more than two ratios are to be decided among. Table 3, Prasert Na Nagara, gives sample size, acceptance regions associated with test cross ratios 1:1, 3:1, and 7:1, probabilities of making wrong decisions when each of the various ratios is true, and dividing lines expressed as ratios for use with any sample size. Table 4 (ibid.) gives the same information for F_2 ratios 9:7, 13:3, and 15:1, etc.

Some of the sample sizes given may be a bit discouraging to you, especially where there are more than two ratios to be decided among. However, they constitute a warning to the person who would base his conclusions on too scant evidence. The tables do enable one to choose a sample size prior to the conduct of the experiment, to give certain assurance against the various possible errors. In addition, one may find them helpful in indicating what type of cross is most efficient in terms of sample size, for testing an hypothesis. Finally, it is possible to use prior information in selecting among ratios to be included among the possible decisions. If no prior information is available, it is permissible to use a small randomly selected portion of the sample to suggest a limited number of hypotheses and then make the test on the remaining data. This is not equivalent to testing hypotheses suggested by the data.

Again, we call to your attention that only a partial solution to the testing of genetic ratios has been presented here. Independent assortment of genes has been assumed and the problem of linkage ignored. In addition, one may be testing between hypotheses 9:7 and 13:3 when the true ratio is 9:6:1 because no one has ever observed the last of the three categories.

15.6 Adjusted chi-square. While the chi-square criterion for testing an hypothesis about a ratio may be relatively unsatisfactory when genetic ratios are involved, it may be quite adequate for many non-genetic problems. Thss, it may be required to test an hypothesis such as, that a certain population is split 50:50 in its opinion concerning the use of some commercial product. Here the alternative may simply be that the population is split in opinion in a ratio other than 50:50. The pollster may be satisfied with some fixed assurance that he is not going to reject this hypothesis with more than a certain probability and consequently will be willing to fix his Type I error. The χ^2 criterion is useful here. If his sample size is small, then it is possible to use the chi-square approximation with a modification to improve the accuracy of the usual probability statement and make the test more nearly equivalent to the binomial test. The adjustment is to counter a bias in the cumulative probability and is called adjustment for continuity, the chi-square criterion being based on a continuous distribution. It is available only for chi-square with one d.f. and is unimportant for $n > 200$.

The adjusted value of chi-square is given by decreasing the absolute value of (observed-expected) by one-half. Thus for our example,

$$\chi^2 = \frac{(|35 - 40.5| - .5)^2}{40.5} + \frac{(|46 - 40.5| - .5)^2}{40.5}$$

$$= 1.23.$$

Subtraction of one-half from (observed - expected) results in a lower chi-square, of course. Consequently, is the unadjusted value of chi-square has been calculated and found to be non-significant, then calculation of the adjusted value will not change the conclusion. However, if the unadjusted value is just beyond the 5% or 1% point, it may be worthwhile to calculate the adjusted value in the interests of improving the accuracy of the probability statement. The point to remember is that the 5% and 1% points have been chosen on an intuitive basis rather that obtained on some sound theoretical basis. Consequently, one should not adopt the attitude that such a percentage point has been endowed with a magical property of absolute power in making decisions between or among hypotheses. At best we have a sample, and are making an uncertain inference.

15.7 Pooling samples. Smith (1927, Amer. Jour. Bot. 14:129-146) considered annual versus biennial growth habit and its inheritance in sweet clover. He examined 38 segregating progenies, of which we present data for the first 6 in Table 15.4.

Table 15.4

Segregating progenies of sweet clover, annual versus biennial habit.

	Observed values			Expected values		
Culture	Annual	Biennial	Total	Annual	Biennial	χ^2
4-3	18	6	24	18.00	6.00	00.000
4-11	33	7	40	30.00	10.00	1.200
4-14	38	12	50	37.50	12.50	0.027
4-15	19	5	24	18.00	6.00	0.223
4-16	39	7	46	34.50	11.50	2.348
4-21	30	6	36	27.00	9.00	1.333
Total	177	43	220	165.00	55.00	5.131

First, test the hypothesis of a 3:1 ratio by calculating individual chi-square values for each culture and a chi-square value based on the pooled data. Each value is distributed as chi-square on 1 d.f. and none of the individual culture values is significant. However, it is to be noted that the value for the pooled data, also

a χ^2 on 1 d.f., is approximately at the 5% point. This suggests that if more similar data were available, then chi-square might have been significant; and reminds us that the power of the test, i.e. our ability to detect small deviations from theory, increases with the size of the sample. No alternative hypothesis is suggested here so that the use of Prasert Na Nagara's tables is not appropriate whereas χ^2 is so.

From table 2 (Prasert Na Nagara), it is seen that a sample of size 210 (ours is 220) can be used to distinguish between the ratios 3:1 and 7:1 at the .01 probability level, i.e. the probability of making an incorrect decision is not more than .01 regardless of which is the true ratio. Further, table 2 (Prasert Na Nagara) is based on the binomial distribution whereas the chi-square test is an approximate one requiring a sample size of about 400 for an observed proportion near .8 in order for normal theory to apply (see Table 15.1) The observed ratio for the pooled samples is .80. Recall that χ^2 on one d.f. involves the square of a single normally distributed variate.

Superiority of table 2 (Prasert Na Nagara) over the use of χ^2 in the situation where 3:1 and 7:1 are the only possibilities is indicated by the fact that the table always tells us to make a decision whereas the use of χ^2 with first one and then the other ratio as the null hypothesis will lead us to reject both hypotheses for certain large samples with proportions near .818.

Let us now look at the sum of the individual chi-square values. Independent chi-square values, when unadjusted, are additive and their sum is distributed as chi-square with d.f. equal to the sum of those associated with the individual chi-squares. We obtain $\chi^2 = 5.131$ on 6 d.f. From Table 9.2, Snedecor, you will find that this is not significant, being somewhere near the 50% point. The question is: How shall we interpret this χ^2 value?

The pooled χ^2 contains an accumulation of the information in all the samples. For example, suppose we sample from a population with a 3:1 ratio and obtain a sample with a χ^2 value of 1.642. This is the value of χ^2 on 1 d.f. such that, on the average, 20% of χ^2 values from random samples from this population will be larger than 1.642. If, however, the true ratio is not 3:1, though this is our hypothesis, and our so-called χ^2 values are distributed about a value

near 1.642 instead, then we would expect the sum of the squares for the 20 samples to be about 20 x 1.642 = 32.84 with 20 d.f. This is significant at the 5% point. Thus, pooling the information from many samples in the form of a χ^2 on multiple degrees of freedom leads us to detect differences not ordinarily detectable when small samples are used.

It has already been stated that a pooling of the data into a single sample and the calculation of a χ^2 on 1 d.f. is also a method of pooling the information from many samples. Thus, the question of the relationship between the two methods of pooling data is raised.

It is easy to visualize a situation involving two χ^2 values, each with a single d.f., where the pooled results would appear to be in accord with the theory although the individuals would be at variance with the theory. In other words, samples need not appear to be from the hypothesized population nor from a common population, yet pooling them may give a result not in conflict with this hypothesis. For example, suppose the following results had been observed on data expected to follow a 3:1 genetic ratio:

Table 15.5

	With	Without	Total	χ^2	d.f.
Sample 1	67	13	80	3.27	1
Sample 2	53	27	80	3.27	1
Total	120	40	160	6.53	2

Each sample is near significance at the 5% level; χ^2 on 2 d.f. is significant at the 5% level; yet the pooled data fit the 3:1 hypothesis perfectly.

Now, let us examine the real data of Tables 15.4 and 15.6. Table 15.6 has a column of observed probabilities. No observed probability is greater than .25, the value called for in our theory. Thus, all six of our samples tend in the same direction from the theoretical ratio and strengthen our decision in favor of the alternative hypothesis. On the other hand, the non-real data of Table 15.5 show observed probabilities which deviate equally on either side of the theoretical ratio, each deviating to the point of being almost significant. In the case of the real data, we can say that the samples appear homogeneous whereas the non-real data are

composed of non-homogeneous data. This lack of homogeneity is also called interaction. We shall discuss the term later with data for which it is a more appropriate term. Here, we will simply think of it as having to do with non-homogeneity.

15.7 Measuring interaction. The χ^2 which consists of the sum of the individual values and that which is obtained from the pooled data both accumulate information from the series of experiments. Subtraction of these χ^2's gives a measure of the homogeneity or heterogeneity, the latter usually being called interaction, of the several ratios. For our data, $\chi^2 = 5.131 - 3.491 = 1.640$ and has $6 - 1 = 5$ d.f. associated with it. This is somewhere near the middle of the distribution of χ^2 with 5 d.f. and we have no evidence to indicate that the progeny from the various cultures are heterogeneous. The contrary is true of our hypothetical data. Here all of the χ^2 on 2 d.f. goes into the interaction χ^2 on 1 d.f. Significance is close to the 1% point.

Table 15.6

Culture	Biennial	Total	P(biennial)	Products
4-3	6	24	.250	1.500
4-11	7	40	.175	1.225
4-14	12	50	.240	2.880
4-15	5	24	.208	1.016
4-16	7	46	.152	1.064
4-21	6	36	.167	1.002
Totals:	43	220		8.687

$\bar{p} = 43/220 = .195$ $\quad \bar{p}(43) = 8.385$

Difference = .302

Table 15.6 illustrates an alternative method of calculating the interaction χ^2 which does not involve the calculation of the individual χ^2's. Column 4 gives the observed probability of the characteristic, biennial; column 5 consists of products of columns 2 and 4. The characteristic, annual, could just as well have been used but it is generally more convenient to use the characteristics with the smaller numbers. Interaction χ^2 is calculated by

$$\chi^2 = \frac{8.687 - 8.385}{(.25)(.75)} = 1.611.$$

The answer differs from the preceding one due to rounding errors. This chi-square is calculated under the assumption that the true ratio is 3:1. Later, a method will be given for calculating χ^2 with no assumption about the true ratio.

15.8 More than two classes. Robertson gives data which include the F_2 progeny of a barley cross from F_1 normal green plants. The observed characters are non 2-row vs. 2-row (Vv) and chlorina vs. normal plant color. These data are given in Table 15.7 and it is desired to test the hypothesis of a 9:3:3:1 ratio (normal dihybrid segregation). χ^2 is calculated as

$$\chi^2 = \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = 54.36 \text{ with } 3 \text{ d.f.},$$

and is highly significant. The evidence is against the theoretical ratio of 9:3:3:1.

Table 15.7

	Non 2-row green	2-row green	Non 2-row chlorina	2-row chlorina	
Observed	1178	291	273	156	1898
Expected	1067.6	355.9	355.9	118.6	
$\frac{(\text{Diff})^2}{\text{Expected}}$	11.416	11.835	19.310	11.794	

What was our alternative hypothesis? Simply that the ratio was other than 9:3:3:1. When an alternative hypothesis is specified, we would like a more efficient method of making a decision among the possible alternatives. Although the multinomial distribution is indicated, it is difficult to define regions of acceptance and rejection. Nothing is presentiy available such as we have for two category ratios in Tables 15.2 and 3, although tables are available for a few 3-category ratios.

15.9 Association and independence. Here are two more words used in connection with χ^2 values for p x q tables. The choice of one of these words or interaction or homogeneity is dependent on the type of data collected. The value of χ^2 is not affected by the name we choose to refer to the phenomenon.

Di Raimondo experimentally infected white mice with staphylococcus aureus to check the behavior in vivo of S.a. cultures grown for several generations in a broth enriched with one of four vitamins. He also used two control broths; in one, he grew the S.a. in the simple broth, referred to as standard; in the other, he grew them in the presence of 0.15 U./ml of penicillin, referred to as penicillin. Injection of the bacterial inoculum was carried out in a number of experiments involving 10 or 20 untreated or treated white mice. The

pooled data from the experiments with untreated mice and the two controlled broths are given in Table 15.8. A logical question to ask is whether or not the percentage of dead (or alive) animals differs according to the culture in which the inoculum is grown, i.e., is the percentage dead associated with the culture or independent of it. Alternatively, we ask whether the two sets of data are homogeneous.

Table 15.8

	Observed			Expected		Dev'ns.	
	Alive	Dead	Total	Alive	Dead		
Standard	8	12	20	8.62	11.38	-.62	.62
Penicillin	48	62	110	47.38	62.62	.62	-.62
	56	74	130	56.00	74.00		

Chi-square is

$$\chi^2 \quad \frac{(-.62)^2}{8.62} + \frac{(.62)^2}{11.38} + \frac{(-.62)^2}{62.62} + \frac{(.62)^2}{47.38} = .0927, \text{ 1 d.f.},$$

and is obviously not significant. We conclude that the percentage dead is independent of the type of control broth used to culture the inoculum.

Notice that the signed deviations add to zero in both rows and columns. This is true of p x q tables where χ^2 for interaction is being calculated with no hypotheses about the true ratio. Theoretical or expected values are calculated from marginal totals. These marginal totals yield an estimate of the probability of death (or survival) on the assumption of independence; here the observed probability of death is 74/130 and of survival is 56/130. The fraction of animals inoculated with bacteria from the standard broth is seen to be 20/130. Thus if there is independence, then on the average, we expect (56/130)(20/130)130 = 8.62 animals inoculated with the standard culture to be alive at the end of the experiment. Deviations from this average are assumed to be approximately normally distributed and χ^2 is the criterion for rejecting or accepting the hypothesis of independence. Obviously, this is an approximation since the observed values must be integers. By some cancellation, you will see that the theoretical value is the product of the marginal totals opposite the cell in question divided by the grand total, a simpler form for calculation.

No hypothesis was required about probabilities; estimates were obtained from data. This makes our example different from that involving the sweet clover. For that example, χ^2 for independence with no assumption about the true genetic ratio is 1.9583 with 5 d.f. If you wish to calculate this by the general 2 x n method given with Table 15.6, you will see that the numerators are the same but that the denominator involves $\bar{p}(1 - \bar{p})$ instead of the theoretical proportions.

Independence χ^2 for the 2 x 2 table can be more conveniently calculated by the formula

$$\chi^2 = \frac{(ad - bc)^2 n}{(a + b)(c+d)(a+c)(b+d)}$$

for data in a table of the form

a	b
c	d.

The numerator is seen to be the difference of the products of the elements in the diagonals, squared and multiplied by n, the total. The denominator is the product of the marginal totals. In the above example,

$$\chi^2 = \frac{(8 \times 62 - 12 \times 48)^2 \, 130}{56 \times 74 \times 20 \times 110}$$

$$= \frac{80^2 \, 130}{9,116,800}$$

$$= .0913$$

differing from the previous value due to rounding errors.

Since this is a χ^2 on 1 d.f., an adjustment for continuity with small numbers is appropriate. As before, this consists of subtracting 1/2 from the absolute value of each deviation before using the calculation formula. With the formula just given, this reduces to

$$\chi^2 = \frac{(|ad - bc| - n/2)^2 n}{(a+b)(c+d)(a+c)(b+d)} .$$

15.10 Test of independence in 2 x n tables. All the necessary information for the calculation of χ^2 for testing independence in 2 x n tables has been given. Calculate the expected values and use the formula or proceed as in Table 15.6. The d.f. are n - 1.

As an example, consider Di Raimondo's data dealing with untreated mice injected with bacterial inoculum cultured in a broth enriched with vitamins niacinamide (N.A.), folic acid (F.A.), p-amino-benzoic acid (Paba), and vitamin B_6 as pyridoxin, each in excess of 10 μg/ml. The data are given in Table 15.9. The value of χ^2 is 2.67 with 3 d.f. and is not significant. We conclude that the probability of death does not differ from inoculum to inoculum, at least not by more than can be attributed to random sampling.

Table 15.9

	Observed		
Inocula	Alive	Dead	Total
N.A.	10	30	40
F.A.	9	31	40
Paba	9	41	50
B_6	13	27	40

15.11 Additivity of chi-square. It has already been stated that independent χ^2 quantities are additive. This property is very convenient for extracting information from data. In fact, we have made use of it already with the data of Di Raimondo.

The data of Tables 15.8 and 15.9 are parts of a single experiment conducted over a period of time with untreated mice. These data fall naturally into two parts: one dealing with mice inoculated with bacteria raised in vitamin enriched broth of four types, and another with mice inoculated with bacteria grown in two types of control broth. The data within each of the two natural parts are no more variable than is to be expected within homogeneous material. This suggests the pooling of the data within a part to obtain a better estimate, one based on larger numbers, of the probability of death to be associated within that part of the data. Then, the question is naturally raised as to whether or not the two resulting observed probabilities differ by more than is to be expected due to random sampling from a common population. The question raised here would normally be raised prior to the conduct of the experiment and would not be suggested by the data. However, observe that if the data had proven to be non-homogeneous within groups, then we would hesitate to test the last-mentioned hypothesis because of the difficulty of interpreting any result.

Table 15.10 consists of the pooled data from the two parts of the experiment. Chi-square for interaction is equal to 12.10 on 1 d.f. and is highly significant. The experimenter will find it hard to explain this difference on the basis of chance alone but will carefully consider all aspects, including possible accidents, before attributing the difference to any one cause such as the vitamin enrichment of the broth or to a number of causes.

Table 15.10

Controls	56	74	130
Vitamins	41	129	170
Totals	97	203	300

$$\chi^2 = \frac{(56 \times 129 - 41 \times 74)^2 \times 300}{13 \times 170 \times 97 \times 203} = 12.10 \text{ with 1 d.f.}$$

The sum of the two χ^2's on single d.f. and the one on 3 d.f. add to that on 5 d.f. within less than 1/2 a unit in almost 15, an error of less than 5%, presumably due to rounding errors.

15.12 The r x c table. The formula

$$\chi^2 = \Sigma \left[\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \right]$$

was originally suggested by Karl Pearson for testing the goodness-of-fit of data to a multinomial model. This was prior to any general theory concerning the testing of hypotheses. It is an approximation which is surprisingly good for small sample sizes provided there is more than a single d.f. The r x c table has (r-1)(c-1) d.f. Expected or theoretical values are easily calculated as the product of the marginal totals opposite the cell in question, divided by the total sample size.

For example, Gilby classified 1725 children according to intelligence and apparent family economic level. A condensed classification gave:

	Dull	Intelligent	Very capable
Very well clothed	81	322	233
Well clothed	141	457	153
Poorly clothed	127	163	48

As for other r x c tables, we estimate the probability to be associated with each cell as the product of the estimated probabilities for the corresponding row or column. The use of the product is based on our hypothesis of independence; if there is dependence or interaction,

then our estimated cell probability will be a poor estimate. Probability times sample size is an estimate of the cell expectation or theoretical value. Table 15.11 gives these values.

Table 15.11

	Dull	Intelligent	Very capable	Total
Very well clothed	128.68	347.31	160.01	636
Well clothed	151.94	410.11	188.95	751
Poorly clothed	68.38	184.58	85.04	338
	349	942	434	1725

$$\chi^2 = \frac{(128.68 - 81)^2}{128.68} + \cdots + \frac{(48 - 85.04)^2}{85.04} = 134.70 \text{ with 4 d.f.}$$

χ^2 is obviously highly significant and we hesitate to explain the result as due to chance alone. We conclude that the hypothesis of independence is wrong. An alternative statement of the same conclusion is that the probabilities of being dull, intelligent, and very capable are not independent of the apparent family economic level.

Each of the upper left, upper right, and lower left cells contribute enough to the value of χ^2 to make it significant. Consequently, it is of interest to consider these cells and surmise on them. Where 129 dull children are to be expected in an apparently high family economic level, only 81 are to be found; where one expects 160 very capable children, 233 are to be found; on the other hand, where one expects 68 dull children in an apparently low family economic level, 127 are observed. Some fairly obvious conclusions can be drawn. The table of contributions certainly seems to be very helpful in formulating a new theory but we cannot test such theories with the data that suggest them. This practice leads too easily to examples of how to lie with figures. The figures themselves do not lie.

15.13 Data with more than two variables of classification. Tables containing such data are often referred to as n-way classification tables. A number of things can be done with data such as this.

With three criteria, the most general hypothesis one might test is that all three criteria of classification are independent. Under this hypothesis, estimates of expected cell values are obtained as products of three probabilities, one for each classification for the category involved with the sample size. Each of the three probabilities is estimated from the sample, the estimates being obtained from mar-

ginal totals as in the case of the two-way classification. One may also test the hypothesis that a stated criteria of classification is independent of the other two. Three such hypotheses are available. Hypotheses of this sort are seen to reduce to those of the two-way classification since two of the classifications are essentially treated as one for purposes of testing. Degrees of freedom for the first case are $(p-1)(q-1)(r-1)$ and for the second are of the form $(pq-1)(r-1)$. It is obvious that the three tests of the latter form are not independent. Other hypotheses may be suggested by the experimenter and we shall treat one of these in this section.

With the exception of the 2 x 2 x 2 table, tables of more than two dimensions are a bit hard to treat. While some hypotheses are relatively easy to test, tests involving partitioning of d.f. are not easily treated in a general manner. Such problems are probably best treated at this stage in the development of statistics by the experimenter and the statistician working together.

Consider data arranged in a 2 x 2 x 2 table as indicated below. Neither figure nor table presentation is entirely satisfactory. These data might be the numbers of germinating and non-germinating seeds from two varieties of a crop, with seeds of each crop receiving a pre-treatment or no treatment at all. There are many possible hypotheses. One or several of these hypotheses should have been in the mind of the experimenter and these should be the only ones under consideration.

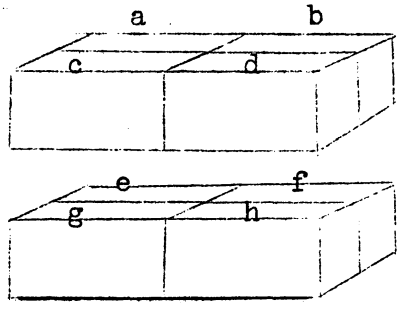


Figure 15.2.

a	b	Variety A
c	d	Variety B
Germ.	Non-G.	

Pre-trt.

e	f
g	h
Germ.	Non-G.

No trt.

For the data of Figure 15.2, we may hypothesize that there is an interaction between variety and germination in both the pre-treated seed and the untreated seed. We wish to test the hypothesis of homogeneity of interaction, i.e. if the interactions differ no more than one would expect due to random sampling from a population in which this sort of interaction was present.

If p_1, p_2, p_3, p_4, p_5, p_6, p_7, and p_8 denote the probabilities of an observation falling in cell a, b, c, d, e, f, g, and h respectively, then the hypothesis of interaction in the pretreated data is that $p_1p_4 = p_2p_3$, and that of interaction in the untreated data is that $p_5p_8 = p_6p_7$. The hypothesis of no second order interaction is that $p_1p_4p_6p_7 = p_2p_3p_5p_8$. There is only one d.f. for this interaction and, consequently, there is only one deviation. This is calculated as the solution of the cubic equation

$$(a+x)(d+x)(f+x)(g+x) = (b-x)(c-x)(e-x)(h-x).$$

The test criterion is

$$\chi^2 = x^2 \sum_{i=1}^{8} \frac{1}{t_i}$$

where t_i is the theoretical value associated with the i-th cell.

As an example, consider the following data of Galton on 78 families. Offspring are classified as light-eyed or not, as to whether or not they had a light-eyed parent, and as to whether or not they had a light-eyed grandparent. The data are

Grandparent		Light		Not	
Parent		Light	Not	Light	Not
Child	Light	1928	552	596	508
	Not	303	395	225	501

The cubic equation is $(1928 + x)(395 + x)(508 + x)(225 + x) = (552 - x)(303 - x)(596 - x)(501 - x)$

and the solution is $x = -30.1$. Chi-square = 16.93 with 1 d.f. Since this is a χ^2 on 1 d.f., the correction for continuity is appropriate. This involves subtraction of .5 from the absolute value of the single deviation, squaring, and dividing by the same sum of squares of reciprocals. Corrected χ^2 is 11.78. The result is still highly significant and it becomes very difficult to explain these data as a result of random sampling from a population where there is no second-order (3-factor) interaction. Alternatively, the two first-order (2-factor) interactions discussed cannot be explained as due to sampling from a single population unless we are prepared to consider this as a most unusual sample. The geneticist will not be surprised at this result.

The test just performed helps answer a question that is often asked. However, before this test is applied, be sure that it involves the question which you wish to ask.

15.14 Continuous data treated as enumeration data. Where a great deal of data are to be or have been collected, it is sometimes decided to treat them as enumeration data in the interests of quick processing. A certain degree of inefficiency will be the result if the data are normally distributed; in this case, the usual techniques for normal data would be valid. However, with sufficient data, the power of tests based on the data in enumeration form is often adequate and the tests are valid regardless of the form of the distribution of the data. Thus, for example, with numerous paired values of wheat yields under two fertilizer treatments, the individual observations might be the signed difference between pairs of yields. If only the sign of the difference is recorded, then the data consist of a sequence of plusses and minuses and an easily tested hypothesis is that these occur with equal frequency subject to the vagaries of random sampling. This hypothesis is seen to be equivalent to that of no treatment difference. Such tests are generally referred to as non-parametric tests.

A measure of association of the variables in an r x c table is given by

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}.$$

It is called the coefficient of contingency and is applicable whether the data arose from discrete or continuous distributions. It is of wide generality because it requires no assumption of underlying normality. Unfortunately, values of C can't be compared unless the tables giving rise to them are of the same dimension.

15.15 Limitations on the analysis of certain four-fold tables. It often occurs that data is available to an individual, yet little or no planning has been involved in the process. Thus, hospitals, doctors, dentists, etc., keep records which it is desirable to examine, yet which may give spurious correlations. Such situations arise under the following circumstances.

In an experiment such as is likely to lead to data in a 2 x 2 table, an experimenter usually has a control and a non-control group, both drawn from a total group of normal plants or animals. On the other hand, in data of the sort just mentioned, one is simply presented with a set of results where it is not known what sort of a group

comprised the original population. Now if each disease has associated with it a probability that its victims will visit a specified doctor for treatment, then it can be shown that diseases occurring independently in the population may appear to be correlated in the group examined even where random fluctuations have been eliminated. This comes as a result of compunding independent probabilities. However, if the selective rate for any particular condition is zero, then the relative incidence of that condition may be validly examined, regardless of selective rates affecting the other groups. For those who wish to examine this phenomenon more closely, we refer you to Berkson, J., "Limitations of the application of 4-fold table analysis to hospital data", Biometrics 2:47-53, 1946.

CHAPTER 18

RAPID PROCEDURES

18.1 Summary. The use of even a minimum electric desk calculator has speeded computing to a point where large collections of data are processed in a short time by a single individual. More complex machinery requiring more highly trained personnel for its operation can process collections of data beyond the scope of even a deluxe desk calculator in the hands of a competent computer using efficient techniques. In either case, it can be said that rapid procedures are in use.

However, this chapter is not meant to deal directly with the use of such equipment. Instead, it is intended to use the chapter to present some of those techniques associated with the so-called distribution-free statistics or non-parametric statistics, not entirely satisfactory terms meant to apply to the parent distribution or distributions. Such techniques are generally easier to apply than those based on assumptions of normality such as have been stressed throughout the book. They are not, in general, as powerful; in other words, more data will generally be required to detect a difference of a stated size than will be required by a test based on normality. On the other hand, they will often be applicable regardless of the form of the parent distribution. This means that they are immediately available to the experimentalist in a variety of situations where he does not feel justified in assuming any specific form of distribution for his observations.

18.2 The sign test. Let A and B represent two varieties of oats in a trial conducted over a wide region involving different soil types, climates and cultivation practices. Past experience merely indicates that the variances of the yields differ over the locations involved. Thus, it may be reasonable to assume a normal distribution but it is not reasonable to assume homogeneity of variances. The validity of the t-test is questioned.

The sign test is based on the signs of the differences between the paired yields. Thus, if x_i and y_i are the yields of varieties A and B respectively in the i-th trial, then if there is no difference

in the yielding ability of the two lines, positive and negative differences should occur with equal frequency. An appropriate test should detect an excess of plus or minus signs beyond that to be expected in random sampling from a binomial population with p = .5. It is to be noted that "no difference" in this case means that the two yield populations have the same median.

T Tables of the binomial distribution can be used to determine e rejection regions, which will usually be two-tailed. At least six pairs of observations are required before a sample will lead to the rejection of the null hypothesis if a 5% level of significance is being used. The test would not ordinarily be used if there were less than 12 paired observations in the sample, and it is preferable to have 20 or more.

Such a test is particularly useful when used in a preliminary testing of the data, when there is a large amount of data, and when assumptions of normality are not considered valid. Paired observations are required with the members of each pair arising under similar conditions. It is not required that the different pairs occur under the same circumstances.

For example. consider the yields of barley grain in a Woburn Rotation Experiment (Cochran, Long Term Agric. Exp. JRSS (B) 6-2-1939) for previous manuring treatment of cotton cake and maize meal. No manures were applied throughout the 12 years of the experiment. Unfortunately, the treatments were not randomized but the yields are obtained from the same plots only once in four years. These yields are given in Table 18.1.

Table 18.1

	Cotton Cake	Maize Meal	Sign (Col. 2 - Col. 3)
1886	207	215	-8
1887	149	156	-7
1888	155	154	+1
1889	141	148	-7
1890	214	193	+21
1891	128	116	+12
1892	142	126	+16
1893	131	125	+6
1894	167	180	-3
1895	76	98	-22
1896	82	80	+2
1897	92	93	-1

There are 6 + signs and 6 - signs. The null hypothesis cannot be rejected on the basis of this evidence.

If the assumptions are that the underlying distribution is normal, then the t-test is valid. In this case, $t = 10/11.8 < 1$ and the same conclusion is drawn. In fact, the probabilities of obtaining a more discrepant result are not very different. On the other hand, suppose the entry in the maize meal column for 1890 read 139 instead of 193. The number of +'s and -'s has not changed but the value of t is now 2.66 and this is significant. The conclusion drawn from use of the sign test has not changed whereas we have an entirely different conclusion if t is relied upon. At worst, the sign test would have changed by only a single sign and there would have been relatively little effect on the conclusions. Looking over the new set of data, the experimenter might well be at a loss in making a decision as to whether or not the value 214 or the value 139 were in error if he relied solely on the data. The distribution-free technique has shown itself to be relatively unaffected by such gross errors.

The sign test supposes that measurements are precise enough to disallow ties. This will not always be the case. Ties are assigned in equal numbers to the + and - categories in most cases. On occasion it may be desirable to leave them out of the test.

18.3 Rank tests of two treatments. To avoid making assumptions of normality when the data do not appear to warrant it, an experimenter may often use rank methods in testing hypotheses. This is possible for two treatments, unpaired or paired. Consider the two laboratories which were sent presumably identical samples of tobacco by a tobacco manufacturer. (Mood, problem 9, p. 283) Nicotine content in milligrams was reported as follows for five determinations: Lab A: 24, 27, 26, 21, and 24; Lab B: 27, 28, 23, 31, and 26. There is no basis for pairing the observations. The question asked is whether or not the two laboratories measure the same thing.

To test this, the 10 observations are ranked. Ties are given the mean rank if in different groups. The rank numbers are added for each lab separately and the smaller rank total noted, in this case 25. Table 18.2 gives the probability of the stated total or a lesser one occurring by chance. The probability levels are similar to those of the F table, i.e. the test as ordinarily used is a one-tailed test. The frequency distribution of rank totals is symmetric

Table 18.2

5% Critical Points of Rank Sums

(smaller) $n_1 \rightarrow$ / $\downarrow n_2$	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4			10											
5		6	11	17										
6		7	12	18	26									
7		7	13	20	27	36								
8	3	8	14	21	29	38	49							
9	3	8	15	22	31	40	51	63						
10	3	9	15	23	32	42	53	65	78					
11	4	9	16	24	34	44	55	68	81	96				
12	4	10	17	26	35	46	58	71	85	99	115			
13	4	10	18	27	37	48	60	73	88	103	119	137		
14	4	11	19	28	38	50	63	76	91	106	123	141	160	
15	4	11	20	29	40	52	65	79	94	110	127	145	164	185
16	4	12	21	31	42	54	67	82	97	114	131	150	169	
17	5	12	21	32	43	56	70	84	100	117	135	154		
18	5	13	22	33	45	58	72	87	103	121	139			
19	5	13	23	34	46	60	74	90	107	124				
20	5	14	24	35	48	62	77	93	110					
21	6	14	25	37	50	64	79	95						
22	6	15	26	38	51	66	82							
23	6	15	27	39	53	68								
24	6	16	28	40	55									
26	7	17	29											
27	7	17												
28	7													

This table is taken from Biometrics 8:1, March 1952, page 37. 1% and 0.1% critical points of rank sums are also available there.

Lab A	Rank	Lab B	Rank
24	3.5	27	7.5
27	7.5	28	9
26	5.5	23	2
21	1	31	10
24	3.5	26	5.5
	21		34

so that two-tailed tests are easily performed. The discrete nature of the distribution means that the tabled probabilities are not exact.

Tests can also be made when the sample sizes are unequal. Here, we calculate the rank total T of the smaller group and the conjugate which is calculated as $n_1(n_1 + n_2 + 1) - T$ where n_1 is the smaller sample size. The smaller number is used in entering the Significance is associated with the smaller numbers since these are obtained when the smaller ranks have a tendency to be associated with a single treatment.

For equal and large sample sizes, rank totals corresponding to .05 and .01 probabilities are sufficiently accurately calculated from the formulas:

$$T_{.05} = \bar{T} - 1.960\sqrt{\frac{N\bar{T}}{6}}$$

$$T_{.01} = \bar{T} - 2.576\sqrt{\frac{N\bar{T}}{6}}$$

where $\bar{T} = 2N(2N + 1)/4$ and N is the number of replicates.

When the treatments are paired, the signed differences are calculated and ranked without regard to sign. These rank numbers are then given the sign of the difference, and rank totals for each sign are then obtained. Reference to Table 18.2a determines significance. This test is against alternatives of the usual analysis of variance type, i.e. the test is one-tailed. For more than 25 reps, values of T for .05 and .01 probability levels can be calculated by the formulas:

$$T_{.05} = \bar{T} - 1.960\sqrt{\frac{(2N + 1)\bar{T}}{6}}$$

$$T_{.01} = \bar{T} - 2.576\sqrt{\frac{(2N + 1)\bar{T}}{6}}$$

where $\bar{T} = N(N + 1)/4$ for N the number of reps.

Table 18.2a

Paired Replicates

Probability (P) of a chance occurrence of rank total of one sign, + or -, whichever is least, equal to ro less than T. T is given in body of table, to nearest whole number. N is number of replicates.

N	P = 0.05	P = 0.02	P = 0.01
6	0	--	--
7	2	0	--
8	4	2	0
9	6	3	2
10	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

The values in this Table were obtained by rounding off values given by J.W. Tukey in Memorandum Report 17, The Simplest Signed Rank Tests, Statistical Research Group, Princeton University, 1949.

This table was taken from Some Rapid Approximate Statistical Procedures, F. Wilcoxon, Am. Cyanamid Co., 1949.

18.4 Rank tests of several treatments in randomized complete block designs. With the use of ranks, the criterion χ_r^2 has been advanced to avoid assumptions of normality. Treatments are ranked within blocks and rank totals for each treatment are then obtained. The value of the criterion is calculated by the formula:

$$\chi_r^2 = \frac{12}{np(p+1)} \Sigma \text{ (rank totals)}^2 - 3n(p+1)$$

where p is the number of treatments and n the number of reps. Probabilities are read from a chi-square table or chart.

The following table shows the percentage success of bolls from flowers for 5 different strains of the cotton plant sown in 1929-30.

Table 18.3

	I		II		III		IV		V		
	%	Rank	%	Rank	%	Rank	%	Rank	%	Rank	Sum
Control	38.2	1	37.7	1	38.9	1	37.9	1	38.2	1	5
Selection A	43.2	2	41.0	2	42.3	2	41.2	2	40.2	2	10
B	46.5	3	45.3	3	45.0	3	45.6	3	44.7	3	15
C	46.8	4	47.4	5	49.3	5	47.1	4	46.5	4	22
D	49.5	5	46.6	4	48.7	4	49.6	5	47.6	5	23
red:	46.5		45.3		45.0		45.6		44.7		

We find

$$\chi_r^2 = \frac{12}{5\,5\,6} (5^2 + 10^2 + 15^2 + 22^2 + 23^2) - 3\,5\,6$$

$$= \frac{12}{150} \quad 1363 \quad - 90$$

$$= \quad 19 \text{ on } p-1 = 4 \text{ d.f.}$$

This is beyond the 1/10th of 1% point as might be expected since the ranks are practically the same in all reps. This example illustrates the use of the χ_r^2 statistic although it is, perhaps, trivial. Percentages are usually considered with suspicion by experimenters till well-considered. Some transformation is often used and a weighted analysis is sometimes resorted to if the denominators are very different. No assumptions about the form of the distribution are required when this rank method is used.

Another test for two-factor experiments is based on medians. Here, the assumption is that the observations have identical distributions except for location. It does not specify that the form of the distribution be normal or of any other stated form.

To calculate the criterion for data from a two-factor experiment with p rows by q columns where it is desired to test the hypothesis of zero row effects, a 2 x p table is first prepared in the following manner: i) find the column medians, ii) find the number of observations in row 1 which exceed their corresponding column median; repeat for the remaining rows, iii) prepare a 2 x p table of the numbers m_i obtained in the previous step and $q - m_i$. The distribution of the chi-square criterion computed for this 2 x p table is very nearly chi-square on p - 1 d.f. unless the marginal totals are quite small.

For example, for the above data the column medians are presented in the last line. For row 1, 38.2 exceeds its column median 46.5, 37.7 exceeds its column median 45.3, etc. The result is the 2.5 table given below.

Table 18.4

Exceed	Not Exceed
0	5
0	5
5	0
5	0
5	0

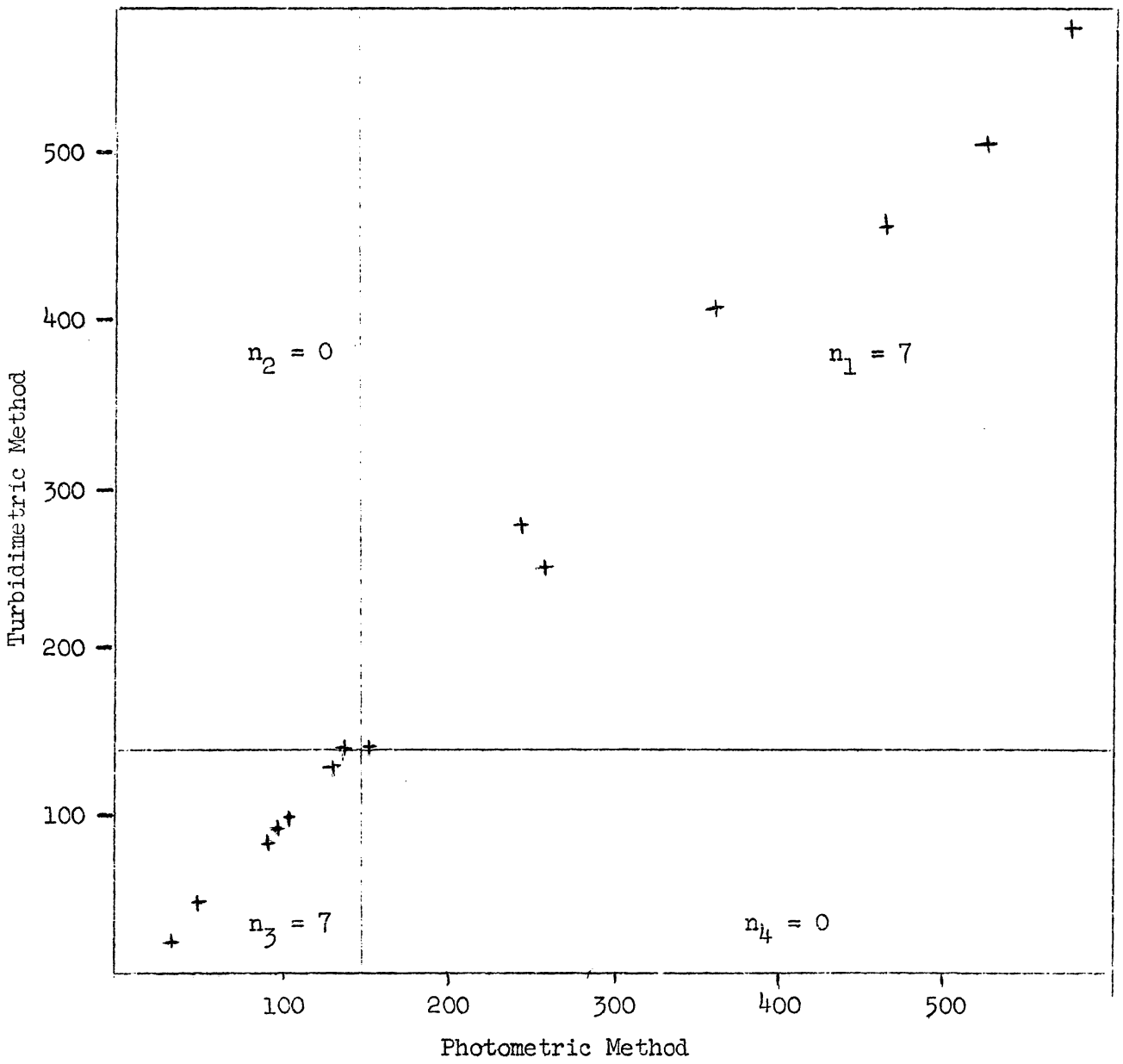
For this table, $\chi^2 =$, has 4 d.f., and is highly significant. Note that any values which may equal the median are counted in the "not exceed" column.

18.5 Measures of association. A fairly common method of quickly obtaining a measure of the linear relation between two variables is to plot the variables on a graph, draw vertical and horizontal lines through the abscissa and ordinate medians respectively, obtain counts of the observations in each of the four quadrants, compute the value of the criterion from these numbers, and use a graph to determine the estimate.

If the quadrants are assigned the numbers n_1, n_2, n_3 and n_4 in a counter-clockwise fashion starting from the upper right corner (+,+) the criterion is computed as $\dfrac{n_1 + n_3}{n_1 + n_2 + n_3 + n_4}$ or $\dfrac{n_2 + n_4}{n_1 + n_2 + n_3 + n_4}$ whichever is larger. The sign of the estimate is obtained by observation of the data. The efficiency of this estimate is low when the data are from a bivariate normal distribution.

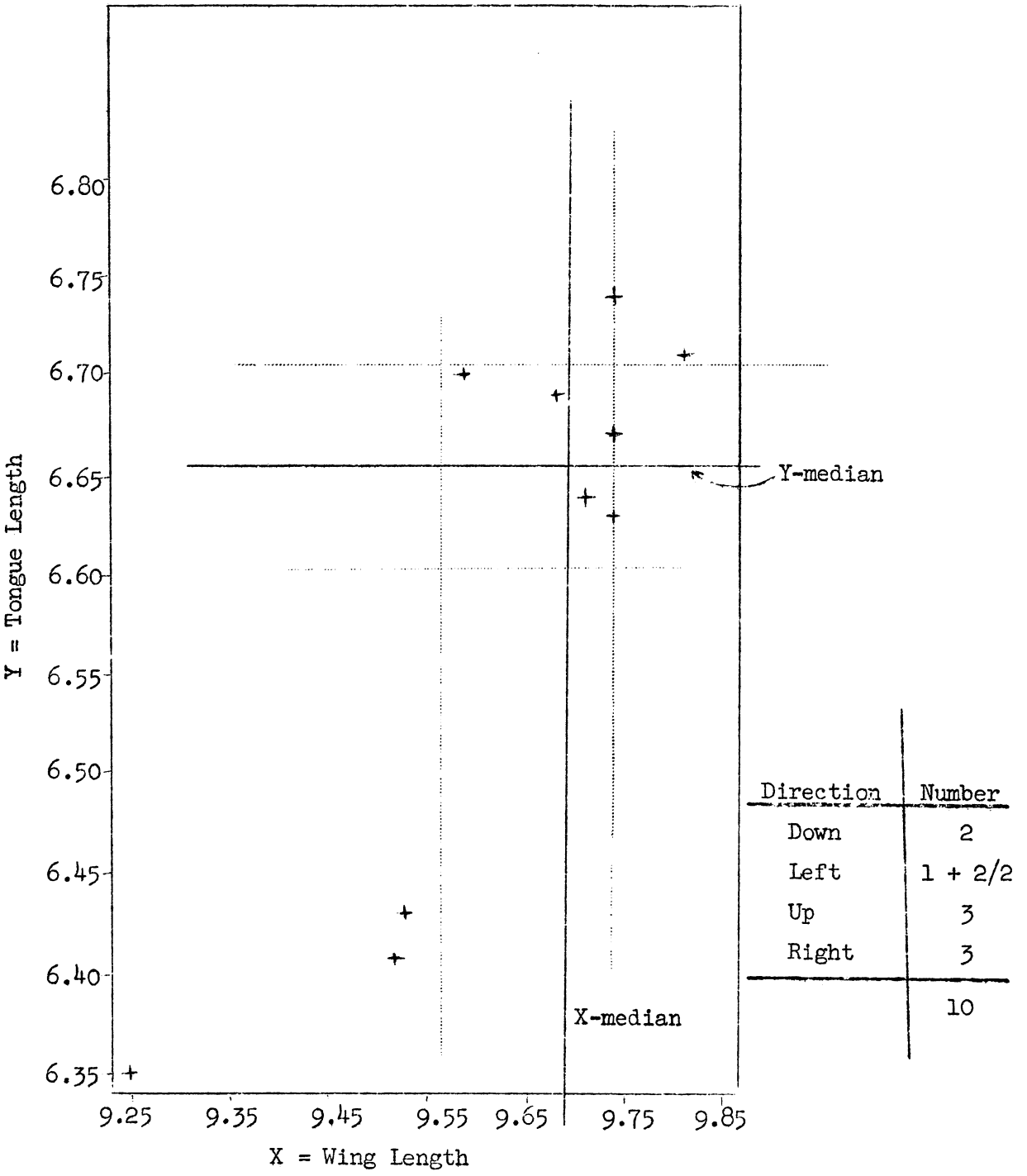
Stanford and English (Use of the Flame Photometer in Rapid Soil Tests for K and Ca, Agronomy Journal, 41-9-1949, p. 446) present the data of Figure 18.1 for a comparison of a rapid photometric and a more time-consuming turbidimetric method for determining the amounts of K present in soils. A relation between the two methods has been established and it would seem desirable to obtain a measure of the linear correlation between the methods. Since all of the data fall in either the (+,+) or the (-,-) quadrant, the criterion must have value 1 as will also the estimate of ρ.

Figure 18.1



To test for association, the quadrant sum test of association is available. To illustrate the procedure, we have selected randomly 10 pairs of observations on wing and tongue length in bees from Grout's (See Snedecor, p. 142) data. Lengths are in millimeters. The data are plotted in Figure 18.2.

Figure 18.2



Direction	Number
Down	2
Left	1 + 2/2
Up	3
Right	3
	10

The first step is to plot the data. Draw a vertical line through the x-median and a horizontal line through the y-median. The resulting four quadrants are assigned signs + for the upper right and lower left and - for the upper left and lower right. From the top of the figure, move down with a sheet of paper or a ruler parallel to the y-median counting the plotted points as they appear till the next point lies across the x-median. This value is +2. Continuing in a clockwise manner, we obtain the values 2, 3, and 3. The algebraic sum called the quadrant sum of 10 is used to enter Table 18.5. This is significant at the 10% level which we regard as giving no evidence against the null hypothesis or as an indication that the experiment is small enough to be inconclusive.

The value 2 resulted from three observations with the same x-value but lying on different sides of the y-median. The procedure is to treat any tied group as if the number of points before crossing the median were

$$\frac{\text{number favorable for inclusion in quadrant sum}}{1 + \text{number unfavorable}}$$

When the number of points is odd and the medians each have a point lying on them, the two points are replaced by a single point with coordinates not involving the medians but taken from the two points that lie on the medians. Proceed with the new point, without the two on the medians, and with the remaining data in the usual manner.

This test is recommended for large amounts of continuous data where preliminary investigations are being performed and speed is desired. It is to be noted that special weight is given to extreme values of the variables. The test is, however, non-parametric.

Table 18.5

Working Significance Levels for Quadrant Sums

Significance Level	Quadrant Sum
10%	± 9
5%	± 11
2%	± 13
1%	± 14-15*

* Use 14 for 14 or more points, 15 for fewer than 14 points.
This table is taken from Some Rapid Approximate Statistical Procedures, Frank Wilcoxon, Am. Cyanamid Co.

Testing Comparisons in the Analysis of Variance

When Snedecor's F-test of the null hypothesis, $H: \mu_1 = \mu_2 = \ldots = \mu_k$ leads to the decision that this hypothesis be rejected, the experimenter is left with the problem of attributing significance to one or more comparisons. Since the experimenter is often in the position of believing that the true hypothesis is contained in the set of alternatives, he must usually face such a problem of explanation.

In many situations, the experimenter has sufficient knowledge of his experimental material and his treatments to plan the testing of certain sub-hypotheses at the same time he is planning his experiment. Examples of this are given in Snedecor:

Pages	Table	Topic
191		partitioning of χ^2
236-7	10.15	samples and sub-samples
267	11.10	use of a standard
273	11.12	regression
277-9	11.14, 11.15	2 types of treatment and their interaction
304	11.13	further partitioning
328	12.8	treatment regression in analysis of covariance

Snedecor gives other examples as well. In theory, each degree of freedom can be associated with a (sum of) square(s) all its own (p. 238). Such partitionings are not unique, as we shall see. In general, exact tests of significance are available for comparisons built into the experimental structure, but no definite probability statements can be made about contrasts suggested by the data (p. 267).

Chapter XV is devoted to a discussion of individual d.f. in relation to the testing of sub-hypotheses. Sets of comparisons may be orthogonal or non-orthogonal (Sections 15.3 and 15.4 define these terms), and, to be tested validly, may require the partitioning of a non-homogeneous error variance.

Such techniques as these serve well in many situations but are inadequate in others. Other techniques are available for many of these situations and we shall examine three of them. However, it must be kept in mind that the greater the experimenter's knowledge of his experimental material, the more meaningful will be the hypotheses he poses for testing, with a consequent increase in the meaningfulness of his inferences.

The least significant difference, l.s.d.: The lsd is defined as that difference between treatment means which would be significant if there were only two treatments. As a result of its definition, it is calculated as: $\sqrt{2/n} \cdot s \cdot t(.05)$ where s is the square root of the error mean square, t(.05) has d.f. as for error mean square, and n is the number of observations in a treatment mean.

It is still not uncommon to see an lsd presented with a set of treatment means and the author may even use it to compare all possible treatment differences. This procedure is invalid. When there are no real treatment differences, a comparison of the highest and lowest means will be declared significant by an lsd about 13% of the time for 3 treatments, 40% of the time for 6 treatments, and 90% of the time for 20 treatments. (Note that this is not equivalent to testing two treatments whose comparison was planned with the design of the experiment and which turned out to be the highest and lowest.)

The trouble with the lsd would seem to be this: when the null hypothesis is true, the error rate is probably the stated 5% on the average, but treatments which rank adjacent to one another are not declared significant as often as 5% of the time whereas treatments with widely differing ranks will be declared significant more than 5% of the time. The error rate for a comparison by lsd of two means, with a given number between, is not generally known. See Snedecor, section 15.5, for a range test.

The gap-straggler-variance test: (see J. W. Tukey, Comparing individual means in the analysis of variance, Biometrics 5:2:1949, p. 99-114.) This test was a step in the right direction but did lead to certain anomalous results. Its author now regards it as obsolete. The test consisted of:

i) Gap test, i.e. ranking the means and comparing adjacent ones by, e.g., an lsd. This locates gaps among the means such that it will be conservative to say that all means on one side of a gap are significantly different from those on the other.

ii) Straggler test. In any group of 3 or more means resulting from step i), test the difference between the group mean and the most straggling mean. The test procedure was given.

iii) Variance test. An F-test for any (sub)group of 3 or more left after the previous steps, test the homogeneity of the remaining groups.

Such a procedure as this is obviously result-guided. As a result, the author guessed that for a tabulated 5% level for falsely declaring significance, the true level was between 6% and 8%.

The New Multiple Range Test: This is the second and simpler test of two proposed by David B. Duncan of Virginia Polytechnic Institute.

The technique permits us to test differences between pairs of means in the analysis of variance. It could well be applied to any sub-set of means chosen prior to the conduct of the experiment as a reasonable set to compare, or to the full set if there is no reason to pose any sub-hypotheses for testing. Let us apply this technique to the data of Table 10.3, Snedecor, which present a set of means with an F-value beyond the 1% point.

Analysis of Variance

Source	d.f.	M. Sq.	F
Fats	7	503.9	3.56**
Mixes within	40	141.6	
Total	47		

$$s_{\bar{x}} = \sqrt{\frac{141.6}{6}} = 4.86$$

Ranked Treatment Means

Treatment	7	8	5	1	6	2	3	4
Mean	161	162	165	172	176	178	182	185

The analysis of variance, $s_{\bar{x}}$, and the ranked treatment means are given above. It is customary to rank the means in increasing order from left to right; however, the same conclusions will be drawn if ranking is in the other order. The problem is to decide which of the 28 differences 4-3, 4-2, 4-6, ..., 8-7 between the 8 means, considered a pair at a time, are significant.

For this, enter table I, a table of significant ranges for a 5% level test at row n_2 = 40 d.f. (we shall have to interpolate) and extract the significant ranges for samples of sizes p = 2, 3, ..., 8. You will get approximately 2.87, 3.02, 3.10, 3.18, 3.23, 3.26, and 3.30. The significant ranges are each multiplied by the standard error, $s_{\bar{x}}$ = 4.86 to form least significant ranges. We have:

i) Least significant ranges

(2)	(3)	(4)	(5)	(6)	(7)	(8)
13.95	14.68	15.07	15.45	15.70	15.84	16.04

ii) Results

161	162	165	172	176	178	182	185

Note: Any two means not underscored by the same line are significantly different. Any two means underscored by the same line are not significantly different.

Notice that the spacing of means is roughly proportional to their numerical differences.

We test the differences: largest minus smallest, largest minus second smallest, ..., largest minus second largest, second largest minus smallest, ..., second smallest minus smallest.

> With one exception, each difference is significant if it exceeds the corresponding least significant range; otherwise it is not significant. Exception: no difference between two means can be declared significant if the two means concerned are both contained in a subset (or the whole set if necessary) with a non-significant range.

The exception to the rule suggests that, as soon as a non-significant difference is found, the two means and the intervening ones be underscored. Further testing of such a subset should not be performed.

The details of the test are:

i) $185 - 161 = 24 > 16.04$; significant

ii) $185 - 162 = 23 > 15.84$; "

iii) $185 - 165 = 20 > 15.70$; "

iv) $185 - 172 = 13 < 15.45$; not significant.

Underline means 172 to 185.

v) $182 - 161 = 21 > 15.84$; significant

vi) $182 - 165 = 17 > 15.70$; "

vii) $182 - 172 = 10 < 15.45$; not significant.

Underline means 172 to 182.

viii) $178 - 161 = 17 > 15.70$; significant

ix) $178 - 162 = 16 > 15.45$; "

x) $178 - 165 = 13 < 15.07$; not significant.

Underline means 165 to 178.

xi) $176 - 161 = 15 < 15.45$; not significant.

Underline means 161 to 176.

Shortcut (when number of means is large): Subtract the least significant range for 8 (in our example) means from the highest mean: 185 - 16.04 ≈ 169. Since 161, 162 and 165 are all less than this value, they are declared to be significantly different from 185. This is so because the significant ranges decrease as the number in the subset decreases. This is seen to cover the first three steps of our example. The idea is repeated with possible elimination of several steps at a time.

The above test is one of several presently available to the experimenter. Definition of type I error would not appear to be the same for all tests nor would the hypotheses being tested seem to be equivalent. If one is not too frightened of declaring a difference to be significant when it is not and wishes to be fairly confident of detecting a difference when it is present, the preceding test would appear to be reasonable.

- -

REFERENCES

Duncan, D. B. Significance tests for differences between ranked treatments in the analysis of variance. Va. Polytechnic Inst. Tech. Report 3, mimeo, June 1953.

Duncan, D. B. Multiple range and multiple F tests. V.P.I. Tech. Report 6, mimeo, September 1953.

Keuls, M. The use of the "studentized range" in connection with an analysis of variance. Euphytica 1:112-122, 1952.

Scheffé, H. A method for judging all contrasts in the analysis of variance. Biometrika 40:87-104, 1953.

Tukey, J. W. The problem of multiple comparisons. Ditto, Princeton Univ., 396 pp., 1953.

Table I. Significant Ranges for a 5% Level New[1] Multiple Range Test

n_2 \ p	2	3	4	5	6	8	10	14	20	50	100
4	3.93	4.01	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02
5	3.64	3.74	3.79	3.83	3.83	3.83	3.83	3.83	3.83	3.83	3.83
6	3.46	3.58	3.64	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.68
8	3.26	3.39	3.47	3.52	3.55	3.56	3.56	3.56	3.56	3.56	3.56
10	3.15	3.29	3.37	3.43	3.46	3.47	3.47	3.47	3.48	3.48	3.48
12	3.08	3.23	3.33	3.36	3.40	3.44	3.46	3.46	3.48	3.48	3.48
14	3.03	3.18	3.27	3.33	3.37	3.41	3.44	3.46	3.47	3.47	3.47
16	3.00	3.15	3.23	3.30	3.34	3.39	3.43	3.45	3.47	3.47	3.47
18	2.97	3.12	3.21	3.27	3.32	3.37	3.41	3.45	3.47	3.47	3.47
20	2.95	3.10	3.18	3.25	3.30	3.36	3.40	3.44	3.47	3.47	3.47
24	2.92	3.07	3.15	3.22	3.28	3.34	3.38	3.44	3.47	3.47	3.47
30	2.89	3.04	3.12	3.20	3.25	3.32	3.37	3.43	3.47	3.47	3.47
60	2.83	2.98	3.08	3.14	3.20	3.28	3.33	3.40	3.47	3.48	3.48
100	2.80	2.95	3.05	3.12	3.18	3.26	3.32	3.40	3.47	3.53	3.53
∞	2.77	2.92	3.02	3.09	3.15	3.23	3.29	3.38	3.47	3.61	3.67

[1] Using special protection levels based on degrees of freedom.